SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 1–20, January 1998

001

ON THE LEAST SQUARES APPROXIMATION OF SYMMETRIC-DEFINITE PENCILS SUBJECT TO GENERALIZED SPECTRAL CONSTRAINTS*

MOODY T. CHU† AND QUANLIN GUO†

Abstract. A general framework for the least squares approximation of symmetric-definite pencils subject to generalized eigenvalues constraints is developed in this paper. This approach can be adapted to different applications, including the inverse eigenvalue problem. The idea is based on the observation that a natural parameterization for the set of symmetric-definite pencils with the same generalized eigenvalues is readily available. In terms of these parameters, descent flows on the isospectral surface aimed at reducing the distance to matrices of the desired structure can be derived. These flows can be designed to carry certain other interesting properties and may be integrated numerically.

Key words. matrix pencil, generalized eigenvalue, symmetric-definite pencil, inverse problem, least squares, descent method, isospectral surface

AMS subject classifications. 65F15, 15A04, 65K10, 49D07

PII. S0895479895285135

1. Introduction.

Let A and B be two square matrices of size n. A matrix pencil of A and B is a family of matrices $A - \lambda B$, parameterized by $\lambda \in \mathbb{C}$. Elements in the set $\sigma(A, B)$ defined by

$$\sigma(A, B) := \{z \in \mathbb{C} | \det(A - zB) = 0\} \tag{1}$$

are called the *generalized eigenvalues* of the pencil. It is easy to see that there are n generalized eigenvalues if and only if $\text{rank}(B) = n$. If B is rank deficient, then $\sigma(A, B)$ may be finite, empty, or infinite. Generalized eigenvalues are preserved under equivalence transformations, i.e., $\sigma(A, B) = \sigma(Y^H AX, Y^H BX)$, provided X and Y are nonsingular matrices and Y^H denotes the conjugate transpose of Y.

In this paper we shall limit our discussion to $\mathbb{R}^{n\times n}$, the Euclidean space of all $n \times n$ real-valued matrices equipped with the Frobenius inner product

$$\langle X, Y\rangle := \sum_{i,j} x_{ij} y_{ij}. \tag{2}$$

For convenience, we also introduce the notation $G(n)$ and $s(n)$ representing, respectively, the general linear group of all nonsingular matrices and the linear subspace of all symmetric matrices in $\mathbb{R}^{n\times n}$. It is frequently the case in practice, and will be assumed henceforth, that A is symmetric and B is symmetric and positive definite. Pencils of this variety are referred to as *symmetric-definite pencils* [7]. For convenience, the corresponding pair of matrices are referred to as a *symmetric-definite pair*.

Obviously $A - \lambda B$ is symmetric definite if and only if $P^T AP - \lambda P^T BP$ is symmetric definite for all $P \in G(n)$. This congruence transformation naturally delineates

* Received by the editors December 14, 1995; accepted for publication (in revised form) by G. Styan December 6, 1996. The research of the first author was supported in part by National Science Foundation grant DMS-9422280.

http://www.siam.org/journals/simax/19-1/28513.html

† Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (chu@math.ncsu.edu, qguo@eos.ncsu.edu).

a "parameterization" for the set

$$\mathcal{M}(A,B) := \{(P^TAP, P^TBP) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} | P \in G(n)\}. \tag{3}$$

We shall show that $\mathcal{M}(A,B)$, consisting of all symmetric-definite pairs with the same generalized eigenvalues $\sigma(A,B)$, is made up of smooth submanifolds in $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$.

This paper concerns the construction of a symmetric-definite pencil satisfying simultaneously conditions on its structure and spectrum. We cast the problem as a task of finding the shortest distance between the set of structured matrices and the *isospectral* set $\mathcal{M}(A,B)$, where $\sigma(A,B)$ is the prescribed spectrum. The approximation is measured by the Frobenius norm over the product space $s(n) \times s(n)$, so a solution is best in the sense of least squares.

More specifically, let V_i, $i = 1, 2$, denote either a single matrix or an affine subspace in $s(n)$ whose elements, qualified by satisfying certain specified conditions on their structure, are being approximated. Define $\mathcal{P} : s(n) \times s(n) \longrightarrow V_1 \times V_2$ by

$$\mathcal{P}(X,Y) := (\mathcal{P}_1(X), \mathcal{P}_2(Y)), \tag{4}$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ denote, respectively, the projections from $s(n)$ onto V_1 and V_2 with respect to the inner product (2). In case V_i is a singleton, define $\mathcal{P}_i(X) \equiv V_i$. The approximation is considered through the optimization problem

$$\min_{(X,Y)\in\mathcal{M}(A,B)} \frac{1}{2}\|(X,Y) - \mathcal{P}(X,Y)\|^2, \tag{5}$$

i.e., the part of (X,Y) that does not carry the desirable structure is being minimized. We emphasize here that the desirable structure in V_1 can be defined independently of that in V_2.

One important point should be clarified before we move on to the discussion of solving (5). We mention that there are two constraints, the spectrum and the structure, imposed upon an ideal problem. In practice, it may occur that one of the two constraints should be more critical than the other due to, for example, the physical realizability. On the other hand, there are also situations where one constraint could be more relaxed than the other due to, for example, the physical uncertainty. Structural constraint usually is imposed due to the physical realizability. Spectral constraint often carries some physical uncertainty. In reality, it is often difficult to maintain both the spectral constraint and the structural constraint concurrently. When these constraints cannot be satisfied simultaneously, a least squares solution becomes the next best thing we can hope for. Depending upon which constraint is to be enforced explicitly, we would have different ways of defining a least squares approximation. The situation in (5) is such that while the pair of matrices (X,Y) vary among the isospectral surface $\mathcal{M}(A,B)$ and hence keep the spectrum $\sigma(A,B)$, the discrepancy between (X,Y) and the desirable structure is minimized. Another situation, which is not addressed in this paper, is to seek a symmetric-definite pair of matrices (X,Y) in the space $V_1 \times V_2$ (and hence the structure is maintained) so that the discrepancy between the two sets $\sigma(X,Y)$ and $\sigma(A,B)$ is minimized. At first glance, these two situations appear to be quite different. In particular, a parameterization for symmetric-definite pairs of matrices with structure specified by V_1 and V_2 is difficult, if not impossible, to obtain. However, it is remarkable that in certain special circumstances these two seemingly unrelated problems can be shown to be equivalent. One such case is the inverse ordinary eigenvalue problem that has already been discussed in [2]. In this paper, we shall focus on (5) only.

The choices of V_i in the setup make the problem (5) quite versatile in application. We mention three immediate applications below. We shall come back in a later part of this paper to explain more specifically how these problems can be solved by our technique.

Problem 1. Given a symmetric-definite pair of matrices $(\tilde{A}, \tilde{B})$ and real numbers $\lambda_1, \ldots, \lambda_n$, find the least squares approximation (X, Y) to $(\tilde{A}, \tilde{B})$ such that (X, Y) is still symmetric definite but $\sigma(X, Y) = \{\lambda_1, \ldots, \lambda_n\}$.

A question that resembles Problem 1 but in the context of ordinary eigenvalue problems, i.e., when $Y \equiv \tilde{B} = I$, can be answered by the Wielandt–Hoffman theorem [4, 10]. For generalized eigenvalue problems, however, the perturbation theory is much more complicated. See, for example, [17, Chapter VI, section 3]. Our approach, by taking $A = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $B = I$ in the definition of the isospectral surface $\mathcal{M}(A, B)$, and $V_1 \equiv \tilde{A}$ and $V_2 \equiv \tilde{B}$ in the definition of the projection $\mathcal{P}$, offers an interesting and easy way to solve Problem 1.

Problem 2. Given a symmetric-definite pencil $A - \lambda B$, find all its generalized eigenvalues.

Among the well-known numerical methods for the symmetric (ordinary) eigenvalue problem, one idea of Jacobi is to systematically reduce the norm of off-diagonal elements. A similar idea can be applied to Problem 2 if we take V_1 and V_2 to be the subspace of all diagonal matrices. In this way, the minimization in (5) amounts to reducing the off-diagonal elements of both X and Y simultaneously by congruence transformation. We shall see that a simple analysis on the stationary points of (5) re-establishes the well-known fact that any symmetric-definite pair can be simultaneously diagonalized.

Problem 3. Given a symmetric-definite pair $(\tilde{A}, \tilde{B})$ and values $\lambda_1, \ldots, \lambda_n$, find a diagonal matrix D so that $\sigma(\tilde{A} + D, \tilde{B}) = \{\lambda_1, \ldots, \lambda_n\}$.

Generalized eigenvalue problems arise, for example, when a Sturm–Liouville problem is discretized by high-order implicit finite difference schemes [14]. An inverse problem, such as Problem 3, is then to reconstruct a certain physical parameter from the natural frequencies. Research on inverse (ordinary) eigenvalue problems has been extensive and fruitful. See, for example, [8] and the references contained therein. Obviously, if $\tilde{B} = \tilde{L}\tilde{L}^T$ is the Cholesky decomposition of $\tilde{B}$, then Problem 3 can be reformulated as finding D such that $\sigma(\tilde{L}^{-1}(\tilde{A} + D)\tilde{L}^{-T}, I) = \{\lambda_1, \ldots, \lambda_n\}$, which becomes an inverse ordinary eigenvalue problem. On the other hand, we may choose, among several options to be discussed in what follows, V_1 to be the affine subspace of $\tilde{A}$ plus all diagonal matrices, $V_2 \equiv \tilde{B}$, $A = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$, and $B = I$. Our approach avoids the inversion of any matrix and guarantees a least squares solution even if an exact solution does not exist.

The multiplicative inverse eigenvalue problem is another important class of problem in applications. The question centers around finding a diagonal matrix D^{-1} so that the "preconditioned" matrix $D^{-1}M$ possesses a specialized spectrum. A multiplicative inverse eigenvalue problem can be formulated as an inverse generalized eigenvalue problem $M - \lambda D$ in a setting similar to Problem 3 except the first entry M is held constant instead.

Solving (5) by standard techniques for constrained optimization problems is not easy because of the matrix structure involved. The main point of this paper is to cultivate descent flows on $\mathcal{M}(A, B)$ for solving (5) in general. Our approach offers a new channel for tackling generalized spectrally constrained problems. The scheme of following flows in the open set $G(n)$ has a similar spirit of an interior-point method

[9, 19], an area that has attracted enormous attention in recent years. However, our methods differ from the traditional interior-point methods in several aspects: neither our objective function nor our feasible set is convex [1, 13, 18], and for most of our flows the dynamics is directed by the objective value rather than the penalty function [20]. We shall comment on this connection again at the end of Example 1 in section 5.

This paper is organized as follows: In section 2 we begin to study the geometry of the isospectral set $\mathcal{M}(A, B)$. We shall show by the algebraic curve theory that $\mathcal{M}(A, B)$ is a union of smooth manifolds. We even can count its dimension in the generic case. In section 3 we outline a framework from which specific differential equations can be designed based on needs or circumstances. The differential equations produce descent flows for (5). Our approach is flexible, yet it offers some theoretical insights as well as ready-made numerical algorithms. In an earlier paper [4], projected gradient flows were derived for least squares approximations with ordinary spectral constraints. Our development here is similar, except that no projection of the gradient is needed this time because $G(n)$ itself is an open set in $\mathbb{R}^{n\times n}$. On the other hand, it will become clear in our study that in order for a flow to maintain a certain additional property, such as being defined on $\mathcal{M}(A, B)$ without reference to its parameterization, the descent direction somehow has to be a modification of the gradient. This point will become manifest in section 3. We highlight some specific applications in section 4. Finally, in section 5 we report some numerical experiments.

2. Isospectral surface. When we refer to flows we mean integral curves of a differential system. To define flows on the set $\mathcal{M}(A, B)$, we have to be certain first of all that $\mathcal{M}(A, B)$ is made of smooth entities. Toward this, we establish two results in this section concerning the topology of $\mathcal{M}(A, B)$.

THEOREM 2.1. *Given any symmetric-definite pair of matrices (A, B), the set $\mathcal{M}(A, B)$ consists of all symmetric-definite pairs with generalized eigenvalues $\sigma(A, B)$.*

Proof. It is clear that if $(X, Y) \in \mathcal{M}(A, B)$, then $X-\lambda Y$ is symmetric definite and $\sigma(X, Y) = \sigma(A, B)$. It is known that any symmetric-definite pencil can be simultaneously diagonalized by congruence transformations. Therefore, if a symmetric-definite pencil $X - \lambda Y$ has the same generalized spectrum $\sigma(A, B)$, then $X - \lambda Y$ is congruent to $\mathrm{diag}(\sigma(A, B)) - \lambda I$ and hence to $A - \lambda B$. This proves the assertion. ∎

The definition (3) may be thought of as an algebraic way to parameterize the set $\mathcal{M}(A, B)$. Note that the parameters come from $G(n)$ which is an open set in $\mathbb{R}^{n\times n}$. The parameterization implies, therefore, that $\mathcal{M}(A, B)$ can be a geometric entity of dimension *at most* n^2. More precisely, we have the following theorem.

THEOREM 2.2. *For any given symmetric-definite pair (A, B), $\mathcal{M}(A, B)$ is a disjoint union of smooth manifolds, each of which has only a finite number of components and has dimension at most n^2 in $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n}$.*

Proof. Consider the vector $c(X, Y) := [c_1(X, Y), \ldots, c_n(X, Y)]^T$ whose components are defined by the coefficients in the polynomial

$$\det(X - zY) = (-1)^n \det(Y) z^n + \sum_{i=0}^{n-1} c_{n-i}(X, Y) z^i.$$

Clearly each $c_k(X, Y)$ is a polynomial in the entries of X and Y. Suppose $\sigma(A, B) = \{\lambda_1, \ldots, \lambda_n\}$. Consider the algebraic variety

$$\mathcal{V}(\lambda_1, \ldots, \lambda_n) := \{(X, Y) \in s(n) \times s(n) | c(X, Y) = (-1)^n \det(Y)\gamma\}, \tag{6}$$

where $\gamma := [\gamma_1, \ldots, \gamma_n]^T$ with $\gamma_k := (-1)^k \sum_{i_1 < \cdots < i_k} \lambda_{i_1} \ldots \lambda_{i_k}$. It follows from Whitney's stratification theorem [12, Theorems 2.3 and 2.4] that $\mathcal{V}(\lambda_1, \ldots, \lambda_n)$ can be expressed as a finite disjoint union of smooth manifolds, each of which has only a finite number of components. Observe that

$$\mathcal{M}(A, B) = \mathcal{V}(\lambda_1, \ldots, \lambda_n) \cap (s(n) \times \mathcal{C}(n)),$$

where $\mathcal{C}(n)$ is the cone of symmetric and positive definite matrices in $\mathbb{R}^{n \times n}$. Since $\mathcal{C}(n)$ obviously is a submanifold in $s(n)$, the assertion follows. □

The gauge n^2 of the dimension is not necessarily an overestimate. We can maintain a little bit more precision on the dimensions of submanifolds involved in Theorem 2.2. A somewhat related discussion can be found in [6]. Let

$$\rho := \max_{(X,Y) \in \mathcal{V}(\lambda_1, \ldots, \lambda_n)} \operatorname{rank} \left[\frac{\partial c}{\partial (X, Y)} \right].$$

Define

$$\mathcal{N}(\lambda_1, \ldots, \lambda_n) := \left\{ (X, Y) \in \mathcal{V}(\lambda_1, \ldots, \lambda_n) | \operatorname{rank} \left[\frac{\partial c}{\partial (X, Y)} \right] < \rho \right\}. \tag{7}$$

Whitney's theorem affirms that $\mathcal{V}(\lambda_1, \ldots, \lambda_n) - \mathcal{N}(\lambda_1, \ldots, \lambda_n)$ is a smooth manifold of dimension $n(n+1) - \rho$. Furthermore, because the rank deficient condition in (7) imposes extra polynomial equations on (X, Y), the set $\mathcal{N}(\lambda_1, \ldots, \lambda_n)$ itself, if not empty, is a union of manifolds with lower dimensions. It follows that $\mathcal{V}(\lambda_1, \ldots, \lambda_n) - \mathcal{N}(\lambda_1, \ldots, \lambda_n)$ is the largest manifold component of $\mathcal{V}(\lambda_1, \ldots, \lambda_n)$ in the sense that $\mathcal{N}(\lambda_1, \ldots, \lambda_n)$ is nowhere dense and has measure zero relative to $\mathcal{V}(\lambda_1, \ldots, \lambda_n)$. Observe that $n(n+1)$ unknowns and n equations are involved in (6), so it must be that $\rho < n$. It follows that the dimension of $\mathcal{V}(\lambda_1, \ldots, \lambda_n) - \mathcal{N}(\lambda_1, \ldots, \lambda_n)$ is *at least* n^2. Together with Theorem 2.2, we conclude that if

$$\mathcal{M}(A, B) \cap \mathcal{N}(\lambda_1, \ldots, \lambda_n) = \emptyset, \tag{8}$$

then $\mathcal{M}(A, B)$ is a smooth manifold of dimension exactly n^2. Sard's theorem [11] guarantees that for almost all choices of (A, B), the condition (8) holds. In particular, it can be shown that (8) holds if (A, B) has distinct generalized eigenvalues. The above result on the parameterization and dimensionality for isospectral symmetric-definite pairs of matrices seems to be known the first time. Though the result may not appear too surprising, the way it is obtained by utilizing Whitney's theorem is of interest in its own right.

Before we move on to describe flows on $\mathcal{M}(A, B)$ we stress that for our application it is not essential whether the set $\mathcal{M}(A, B)$ itself is a one-piece manifold. The differentiable flows that will be defined later automatically stay on smooth components of $\mathcal{M}(A, B)$.

We conclude this section by one example showing that the inverse eigenvalue problems for matrix pencils could be quite intricate. We show that in special circumstances $\mathcal{M}(A, B)$ may be a proper subset of $\mathcal{N}(\lambda_1, \ldots, \lambda_n)$. Consider the case when $n = 2$, $A = 0$, and $B = I$. Then $\mathcal{M}(A, B) = \{(0, P^T P) | P \in G(n)\}$. Though $G(2)$ has dimension 4, $\mathcal{M}(A, B)$ obviously has dimension 3. It is interesting to note that for a pair $X = (x_{ij})$ and $Y = (y_{ij})$ to be in $\mathcal{V}(0, 0)$, a necessary condition is that the entries satisfy the equations

$$x_{21} = x_{12},$$

$$
\begin{aligned}
y_{21} &= y_{12},\\
x_{11} &= |x_{12}|\frac{\mathrm{sgn}(x_{12})y_{12} \pm \sqrt{y_{12}^2 - y_{11}y_{22}}}{y_{11}},\\
x_{22} &= \frac{-y_{11}x_{11} + 2x_{12}y_{12}}{y_{22}},
\end{aligned}
$$

provided $y_{11}y_{22} \neq 0$. There are four free parameters in defining $\mathcal{V}(0,0)$. However, if Y is required to be positive definite, then $X = 0$ is the only possible solution.

3. Descent flows. The parameterization (3) provides grounds for maneuver on $\mathcal{M}(A,B)$ to reduce the objective value in (5). In this section, we discuss how to take advantage of this parameterization to formulate descent flows.

We start with working within the parameter space $G(n)$. For convenience, we introduce the abbreviation

$$
\left\{ \begin{array}{l} \alpha_1(P) := P^TAP - \mathcal{P}_1(P^TAP),\\ \alpha_2(P) := P^TBP - \mathcal{P}_2(P^TBP) \end{array} \right. \tag{9}
$$

when the symmetric-definite pair (A,B) is fixed. The objective function in (5) is equivalent to the function $F : G(n) \longrightarrow R$, where

$$
F(P) := \frac{1}{2}\left(\langle \alpha_1(P), \alpha_1(P)\rangle + \langle \alpha_2(P), \alpha_2(P)\rangle\right). \tag{10}
$$

The following result is critical in our development.

THEOREM 3.1. *The gradient ∇F of F is given by*

$$
\nabla F(P) = 2\left\{AP\alpha_1(P) + BP\alpha_2(P)\right\}. \tag{11}
$$

Proof. Observe that the Fréchet derivative of F at P acting on $H \in \mathbb{R}^{n\times n}$ can be calculated as follows:

$$
\begin{aligned}
F'(P)H &= \langle \alpha_1(P), H^TAP - \mathcal{P}_1'(P^TAP)H^TAP + P^TAH - \mathcal{P}_1'(P^TAP)P^TAH\rangle\\
&\quad + \langle \alpha_2(P), H^TBP - \mathcal{P}_2'(P^TBP)H^TBP + P^TBH - \mathcal{P}_2'(P^TBP)P^TBH\rangle\\
&= 2\left\{\langle \alpha_1(P), P^TAH - \mathcal{P}_1'(P^TAP)P^TAH\rangle\right.\\
&\quad \left. + \langle \alpha_2(P), P^TBH - \mathcal{P}_2'(P^TBP)P^TBH\rangle\right\}\\
&= 2\left\{\langle \alpha_1(P), P^TAH\rangle + \langle \alpha_2(P), P^TBH\rangle\right\}\\
&= 2\langle AP\alpha_1(P) + BP\alpha_2(P), H\rangle. \qquad (12)
\end{aligned}
$$

In the above, the second equality is due to the symmetry of the matrices involved. The third equality follows from the fact that the action of $\mathcal{P}_i'$ (at P^TAP and P^TBP, respectively) on any point (P^TAH and P^TBH, specifically) resides in the tangent space of V_i whereas the range of α_i is perpendicular to the tangent space of V_i. The last equality is obtained by utilizing the adjoint property of the Frobenius inner product. It follows from (12) that the gradient ∇F of F may be interpreted as asserted. ∎

Obviously, the differential equation

$$
\dot{P}(t) := -\nabla F(P(t)), \tag{13}
$$

where $\dot{P}$ means the derivative of P with respect to a certain artificial parameter t, defines the steepest descent flow $P(t)$ on $G(n)$ for F. It should be cautioned, however, that the open set $G(n)$ has a *boundary* made up of all $n \times n$ singular matrices. The differential equation (13) alone cannot guarantee that the flow $P(t)$ will stay away from the boundary of singular matrices. The first example in section 5 clearly illustrates this occurrence.

Through the parameterization relationship

$$\begin{cases} X(t) = P(t)^T A P(t), \\ Y(t) = P(t)^T B P(t), \end{cases} \tag{14}$$

each flow in the parameter space $G(n)$ has a corresponding flow on $\mathcal{M}(A, B)$. Related to the flow $P(t)$ defined by (13), for example, is the flow $X(t)$ defined by

$$\begin{aligned} \dot{X} &= -2\left\{\alpha_1(P)P^T A^2 P + \alpha_2(P)P^T BAP \right. \\ &\qquad \left. + P^T A^2 P\alpha_1(P) + P^T ABP\alpha_2(P)\right\} \qquad (15) \\ &= -2\left\{\beta_1(X)X(P^T P)^{-1}X + \beta_2(Y)Y(P^T P)^{-1}X \right. \\ &\qquad \left. + X(P^T P)^{-1}X\beta_1(X) + X(P^T P)^{-1}Y\beta_2(Y)\right\}, \qquad (16) \end{aligned}$$

where we have denoted

$$\begin{cases} \beta_1(X) := \alpha_1(P), \\ \beta_2(Y) := \alpha_2(P) \end{cases} \tag{17}$$

to emphasize the dependence of the system on the variables X and Y. A similar flow $Y(t)$ can also be defined.

Neither (15) nor (16) is useful in that the differential system depends explicitly on the parameterization variable P. That dependence means that to integrate (15) or (16) one must also integrate (13). This is a waste since the parameter flow $P(t)$ needs to be integrated in any case. It perhaps would be more economical to obtain $X(t)$ and $Y(t)$ directly from (14).

Note also that the system (13) defines the steepest descent flow. There are situations when one prefers to relinquish the steepest descent property in exchange for maintaining other attributes. In the following we introduce several other descent flows for this purpose.

We first illustrate a situation where the description of $X(t)$ and $Y(t)$ can be implicit in the parameter P.

COROLLARY 3.2. *The flow defined by*

$$\dot{P} := -\frac{1}{2} P P^T \nabla F(P) \tag{18}$$

is a descent flow.

Proof. Observe that

$$\langle \nabla F(P), -PP^T \nabla F(P) \rangle = -\langle P^T \nabla F(P), P^T \nabla F(P) \rangle \leq 0$$

and that the equality holds only when $\nabla F(P) = 0$. Thus, the differential system (18), though not the steepest one, continues to define a descent flow for F. ∎

Upon substitution, the corresponding flow $(X(t), Y(t))$ on $\mathcal{M}(A, B)$ is defined by the differential system

$$\begin{cases} \dot{X} = -\left((XW)^T + XW\right), \\ \dot{Y} = -\left((YW)^T + YW\right) \end{cases} \tag{19}$$

with

$$W := X\beta_1(X) + Y\beta_2(Y). \tag{20}$$

Note that the differential system (19) is autonomous in X and Y and makes no reference to the variable P. The computation of $P(t)$ as well the troublesome matrix inversion such as $(P^TP)^{-1}$ in (16) are thus avoided.

It is worth noting that the critical points of the differential system (18) are exactly the same as the stationary points of the optimization problem (5), provided that critical point is nonsingular. The optimization problem (5), therefore, can be solved by integrating (19) from a suitable starting point, say $(X(0), Y(0)) = (A, B)$, until a limit point is located.

The simplest case of (19) when $n = 1$ is rather illuminating. Corresponding to a given pair of numbers (A, B) with $B > 0$, the set $\mathcal{M}(A, B) = \{(X, Y) \in \mathbb{R}^2 | X = AP^2, Y = BP^2, P \neq 0\}$ is a half-array that emanates from but does not include the origin in the direction (A, B). In particular, $\mathcal{M}(A, B)$ is an *unbounded open* set. Suppose we want to solve Problem 1 mentioned in section 1. The corresponding differential system of (19) becomes

$$\begin{cases} \dot{X} = -2X(X(X - \tilde{A}) + Y(Y - \tilde{B})), \\ \dot{Y} = -2Y(X(X - \tilde{A}) + Y(Y - \tilde{B})). \end{cases} \tag{21}$$

All critical points of (21) are included in the set

$$\{(X, Y) | X(X - \tilde{A}) + Y(Y - \tilde{B}) = 0\},$$

which is the dotted circle represented in Figure 1. But relative to $\mathcal{M}(A, B)$, where the flow starting from $X(0) = A$ and $Y(0) = B$ resides, only the two critical points

$$(0, 0) \quad \text{and} \quad \left(\frac{C(C\tilde{A} + \tilde{B})}{I + C^2}, \frac{C\tilde{A} + \tilde{B}}{I + C^2} \right) \tag{22}$$

with $C := AB^{-1}$ are most pertinent. Consider the case when the target point $(\tilde{A}, \tilde{B})$ is located in the *lower* half-plane of the line that passes through the origin and is perpendicular to the array $\mathcal{M}(A, B)$ (see the shaded region in Figure 1.) Obviously the shortest distance from $(\tilde{A}, \tilde{B})$ to $\mathcal{M}(A, B)$ is attained only at the origin, but that point does not belong to $\mathcal{M}(A, B)$. Thus, Problem 1 should have *no* true solution in this case. Nonetheless, the flow defined by (21) stays on the half-array and indeed moves toward the origin. In this way, we end up with a *pseudosolution* in the sense that the solution is still a least squares approximation, but that point is not from within $\mathcal{M}(A, B)$. On the other hand, the second critical point (22) in this case is away from the set $\mathcal{M}(A, B)$ by a positive distance and hence can never be realized. We shall refer back to (21) in section 4 for further discussion of a higher-dimension case.

We next mention two more descent flows that possess some additional interesting properties.

COROLLARY 3.3. *The differential equation*

$$\dot{P} := -\frac{1}{2} P \left\{ P^T \nabla F(P) - \nabla F(P)^T P \right\} \tag{23}$$

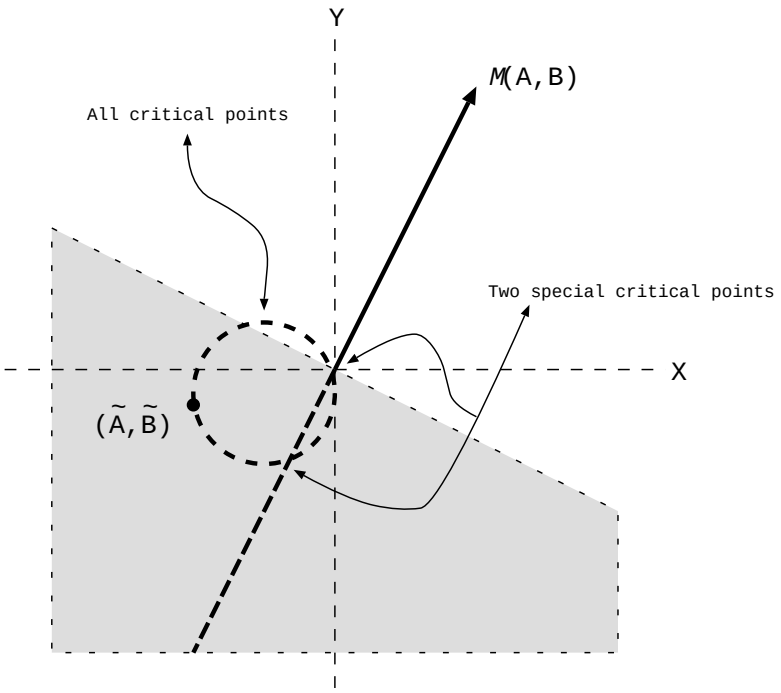


FIG. 1. *Geometry of a pseudosolution.*

defines a descent flow. Furthermore,

$$P(t)P(t)^T \equiv \textit{constant}.$$

Proof. From the fact that

$$\langle \nabla F(P), P\left\{P^T\nabla F(P) - \nabla F(P)^T P\right\}\rangle = \langle P^T\nabla F(P), P^T\nabla F(P) - \nabla F(P)^T P\rangle$$

and the equality that

$$\langle M, M - M^T\rangle = \sum_{j\neq i}(m_{ij} - m_{ji})^2 \geq 0$$

for any square matrix $M = (m_{ij})$, it follows that the flow $P(t)$ enjoys the descent property. Furthermore, because the quantity in the braces of (23) is skew symmetric, it is easy to see that $P\dot{P}^T + \dot{P}P^T = 0$. Thus, $P(t)P(t)^T \equiv P(0)P(0)^T$ for all t. ∎

The corresponding flow on $\mathcal{M}(A,B)$ are integral curves of the double-bracket system:

$$\left\{\begin{array}{rcl} \dot{X} &=& [X,[X,\mathcal{P}_1(X)] + [Y,\mathcal{P}_2(Y)]]\,, \\ \dot{Y} &=& [Y,[X,\mathcal{P}_1(X)] + [Y,\mathcal{P}_2(Y)]]\,, \end{array}\right. \tag{24}$$

where $[X,Y] := XY - YX$ denotes the Lie bracket. Note that the system (24) is autonomous. Note also that if $P(0) = I$ from the beginning, then $P(t)$ remains orthogonal for all t. Our notion here generalizes that of orthogonal similarity transformation discussed in [5].

COROLLARY 3.4. *The differential equation*

$$\dot{P} := -\frac{1}{2}\left\{\nabla F(P)P^T - P\nabla F(P)^T\right\}P \tag{25}$$

is a descent flow. Furthermore,

$$P(t)^T P(t) \equiv \textit{constant}.$$

Proof. The proof is similar to Corollary 3.3. ∎

Although it looks similar to (23), this new system (25) by no means is a trivial alternation (say, by taking the transpose) of (23). In particular, it can be checked by substitution that the corresponding differential equation for $X(t)$ and $Y(t)$ depends explicitly on the variable P in (25), a predicament that does not occur in (24). The system (25) is especially useful for attacking problems where the corresponding flow $Y(t)$ is expected to be constant. Problem 3 is one such instance. We shall be more specific on its application in the next section.

We conclude this section with one remark on the asymptotic behavior of the flows.

THEOREM 3.5. *For all the flows $P(t)$ defined above, the corresponding $(X(t), Y(t))$ converges. Generically, the limit point is a stationary point, possibly on the boundary of $\mathcal{M}(A, B)$, of* (5)*. The nongeneric exception is when the product $P^T\nabla F(P)$ in* (23) *or $\nabla F(P)P^T$ in* (25) *is symmetric at the limit point.*

Proof. Along any solution $(X(t), Y(t))$ the function

$$G(t) := F(P(t)) = \frac{1}{2}\left\{\langle\beta_1(X(t)), \beta_1(X(t))\rangle + \langle\beta_2(Y(t)), \beta_2(Y(t))\rangle\right\} \tag{26}$$

satisfies

$$\dot{G}(t) = \langle\nabla F(P(t)), \dot{P}(t)\rangle \le 0.$$

Furthermore, $\dot{G} = 0$ only when $\nabla F(P) = 0$ or $\dot{P} = 0$. The latter case generically implies also $\nabla F(P) = 0$. Thus, $G(t)$ is monotonically decreasing until a stationary point of (5) is found. ∎

4. Applications. Our differential system approach not only can be used as a convenient algorithm for finding a least squares solution but also offers some theoretical insights into the problem. In this section we explain more specifically how our approach can be applied to solve the three problems described in section 1. We discuss the applications case by case. Further numerical experiments will be reported in section 5.

Application 1. We point out earlier that there is no easy generalization of the Wielandt–Hoffman theorem for Problem 1. To demonstrate the complexity of Problem 1 in general, we consider a very special case when both target matrices $\tilde{A}$ and $\tilde{B}$ are diagonal. Our point of this overly simplified problem is to illustrate how complicated the stationary points for Problem 1 could be. Suppose that the differential equation (19) (which is based on the descent flow (18)) is used to solve the problem from the initial values $X(0) = A = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $Y(0) = B = I$. Recall that the critical points of the differential system are exactly the same as the stationary points of the problem. By construction we know the solution flow $(X(t), Y(t))$ of (19) remains diagonal. The differential system, being uncoupled into n pairs (x_{ii}, y_{ii}), $i = 1, \ldots n$, can be represented exactly by (21) if all symbols there are interpreted as (diagonal) matrices. Observe that the pairs $(x_{ii}(t), y_{ii}(t))$ are independent of each

other and may converge to limit points of different types (see (22)). In particular, some of the pairs, as pointed out earlier, may converge to an infeasible limit point $(0, 0)$. This simple uncoupled system highlights the potential difficulty for general $\tilde{A}$ and $\tilde{B}$ where these events are intertwined together and hence make Problem 1 more complicated. Regardless of this complexity, our differential equation offers an easy-to-use numerical method for solving this type of problem.

Application 2. Using the setup described in Problem 2, i.e., V_1 and V_2 are the subspaces of all diagonal matrices, the first-order optimality condition $\nabla F(P) = 0$ at any stationary point P is equivalent to the equality

$$X(X - \mathrm{diag}(X)) + Y(Y - \mathrm{diag}(Y)) = 0, \tag{27}$$

where X and Y are related to P by (14). It is easy to check that the diagonal elements involved in (27) are given by

$$\sum_{k \neq i} x_{ik}^2 + \sum_{k \neq i} y_{ik}^2 = 0, \qquad i = 1, \ldots, n. \tag{28}$$

That is, (X, Y) is a limit point of the descent flow (19) if and only if both X and Y are diagonal matrices. Our differential equation (19) not only re-establishes the fact that any symmetric-definite pencil can be simultaneously diagonalized but also offers a numerical way to accomplish this.

Application 3. We give a few more details below for Problem 3 since it is of particular interest and importance. The geometry of Problem 3 is sketched in Figure 2 where we use the three-dimensional coordinate axes to represent the triplet $(\text{off-diag}(X), \mathrm{diag}(X), Y)$ for any matrix pair $(X, Y) \in \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n}$. The desirable state, represented by the bold horizontal line in Figure 2, means that $Y = \tilde{B}$ and $\text{off-diag}(X) = \text{off-diag}(\tilde{A})$. The minimization in (5) is equivalent to minimizing the distance between the two points **P** and **Q** in Figure 2 while **P** stays in $\mathcal{M}(A, B)$ (not drawn) and **Q** stays in the desirable state.

The desirable state can be characterized by selecting V_1 to be the affine subspace of $\tilde{A}$ plus all diagonal matrices and $V_2 \equiv \tilde{B}$. To maintain the eigenvalue information, an obvious choice would be letting $A = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ and $B = I$. The projections corresponding to this setup imply that $\beta_1(X) = \text{off-diag}(X - \tilde{A})$ and $\beta_2(Y) = Y - \tilde{B}$. While any of the differential equations we proposed, say (19), is ready for integration, there is a setback in using some of these equations. The resulting solution flow may *stop* at a local minimizer that does not meet the criteria of the desirable state, i.e., the resulting $Y(t)$ is likely to vary in t whereas the second matrix involved in Problem 3 is expected to be constantly $\tilde{B}$.

To remedy the above fault, we may consider using the differential system (25) with initial values

$$P(0) = UL^T, \tag{29}$$

where $\tilde{B} = LL^T$ is the Cholesky decomposition of $\tilde{B}$ and U is an arbitrary orthogonal matrix. Corollary 3.4 guarantees that $Y(t) \equiv \tilde{B}$ and hence $\beta_2(Y(t)) \equiv 0$ for all t. The differential equation (25) becomes

$$\dot{P} = \left[P\alpha_1(P)P^T, \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}\right] P, \tag{30}$$

where $\alpha_1(P) = \text{off-diag}(P^T \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\} P - \tilde{A})$. The Lie bracket operation in (30) is simple because the second operant is a diagonal matrix. The resulting $P(t)$

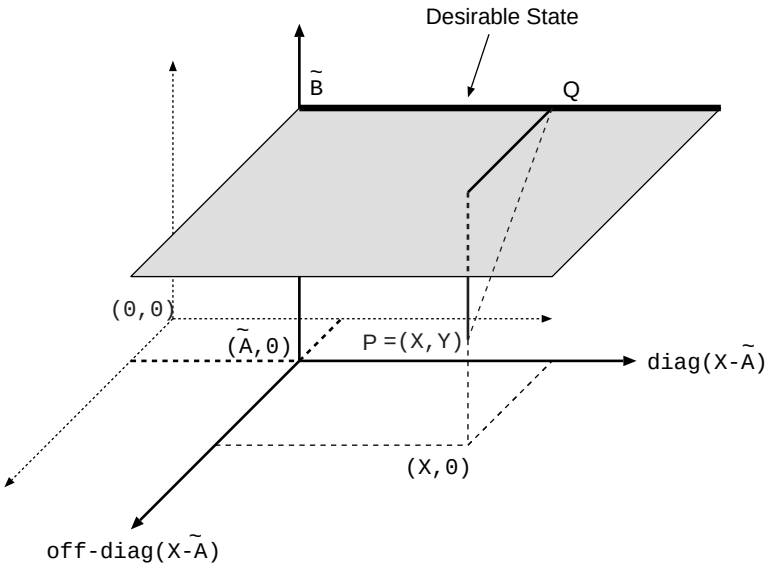


FIG. 2. *Geometry of Problem* 3.

implicitly defines a flow $(X(t), Y(t))$ on the shaded region represented in Figure 2. The flow starts from $(X(0), Y(0)) = (L\text{diag}\{\lambda_1, \ldots, \lambda_n\}L^T, \tilde{B})$ and approximates the set of the desirable state. Once the limit point $P(\infty)$ of (30) is found, the diagonal matrix

$$D := \text{diag}(X(\infty) - \tilde{A}),$$

where $X(\infty) = P(\infty)^T \text{diag}\{\lambda_1, \ldots, \lambda_n\}P(\infty)$, is an optimal solution for Problem 3 in the sense of least squares.

5. Numerical experiment. In this section we report some of our numerical experiments with the proposed methods. At present we are more concerned with the dynamics of the flows than the efficiency of the programs. Thus, we only consider using general-purpose initial value problem software as the integrator. We have experimented with both the FORTRAN code ODE [15] and the MATLAB code ODE SUITE [16]. The results are similar. We shall only report experiments from ODE SUITE since it is easier to manipulate matrix operations and to present the results graphically by MATLAB.

There are two types of solvers, **ode113** and **ode15s**, in the MATLAB ODE SUITE. The code **ode113** is a PECE implementation of Adams–Bashforth–Moulton methods for nonstiff systems. The code **ode15s** is a quasi-constant step size implementation of the Klopfenstein–Shampine family of the numerical differential formulas for stiff systems. The statistics about the cost of integration can be obtained directly from the **odeset** option built in the integrator. More details of these codes can be found in the document [16]. Again we have experimented with both solvers. We discover that when the prescribed eigenvalues do not vary wildly, these two codes perform comparably. But when the ratio of the eigenvalue with the largest magnitude to the smallest gets larger, the **ode15s** becomes faster in terms of CPU time. We think a largely varying spectrum, perhaps, has resulted in a stiff initial value problem.

In our experiments the tolerance for both absolute error and relative error is set at 10^{-12}. This criterion is used to control the accuracy in following the solution path. The high accuracy we required here has little to do with the dynamics of the underlying vector field, and perhaps is not needed in practical application. We examine the output values at time intervals of 1 and 10, and assume that the path has reached an equilibrium point whenever the difference of the Lyapunov's functions (26) at two consecutive output points is less than 10^{-10}. So as to fit the data comfortably in the running text, we report only the case $n = 5$ and display all numbers with five digits.

Example 1. In our first experiment we report one pathological example where the flow $P(t)$ of parameters converges to the boundary of singular matrices, and hence the corresponding least squares problem is solved in an unusual yet interesting way.

Suppose we want to solve the generalized eigenvalue problem, Problem 2, for the pair of matrices

$$A = \begin{bmatrix} 1.0904 & 0.1575 & 0.2394 & 2.5284 & -0.4716 \\ 0.1575 & 0.2913 & -1.0421 & 1.8527 & 0.4591 \\ 0.2394 & -1.0421 & -2.2831 & -0.0859 & -2.2171 \\ 2.5284 & 1.8527 & -0.0859 & -2.5200 & -1.1272 \\ -0.4716 & 0.4591 & -2.2171 & -1.1272 & 1.1959 \end{bmatrix},$$

$$B = \begin{bmatrix} 6.8747 & -1.6174 & -1.3123 & 4.2938 & 0.5968 \\ -1.6174 & 6.8615 & 1.2753 & -2.2454 & -5.3684 \\ -1.3123 & 1.2753 & 2.8018 & 1.2469 & 0.6560 \\ 4.2938 & -2.2454 & 1.2469 & 5.1703 & 1.9403 \\ 0.5968 & -5.3684 & 0.6560 & 1.9403 & 10.6641 \end{bmatrix}$$

by using the steepest descent flow (13) with initial value

$$P(0) = \begin{bmatrix} -0.62735 & -0.04006 & 0.42746 & 0.63529 & 0.13607 \\ -0.41918 & -0.12833 & 0.34523 & -0.51495 & -0.65074 \\ -0.23520 & 0.77311 & -0.42324 & 0.18008 & -0.36799 \\ -0.22678 & 0.49212 & 0.28205 & -0.51204 & 0.60387 \\ -0.56918 & -0.37689 & -0.66288 & -0.19137 & 0.24073 \end{bmatrix}.$$

When our code terminates, suggesting that a convergence has been reached, we discover that

$$P(\infty) = \begin{bmatrix} -0.0243 & -0.1109 & 0.2316 & -0.0000 & 0.1459 \\ -0.1314 & 0.0106 & 0.1922 & -0.0000 & -0.3265 \\ 0.0860 & 0.2712 & 0.0432 & 0.0000 & -0.2058 \\ -0.0279 & 0.2026 & 0.3343 & -0.0000 & 0.2038 \\ -0.0979 & 0.0565 & -0.2861 & -0.0000 & 0.0261 \end{bmatrix}.$$

The fourth column of $P(\infty)$ is in fact as small as

$$\left[-0.20072\times10^{-13},\ -0.13475\times10^{-12},\ 0.94951\times10^{-13},\ -0.32969\times10^{-13},\ -0.97683\times10^{-13}\right]^T,$$

indicating that $P(\infty)$ is nearly singular. Note that this result of near singularity does not contradict condition (28), where we argue that (X, Y) is a stationary point of (5) if and only if both X and Y are diagonal matrices. Indeed, we obtain that

$$X = P(\infty)^T A P(\infty) = \mathrm{diag}\{0.0800, -0.4773, 0.7925, 0.0000, -0.4043\},$$
$$Y = P(\infty)^T B P(\infty) = \mathrm{diag}\{0.0635, 0.6128, 2.4657, 0.0000, 2.1823\}.$$

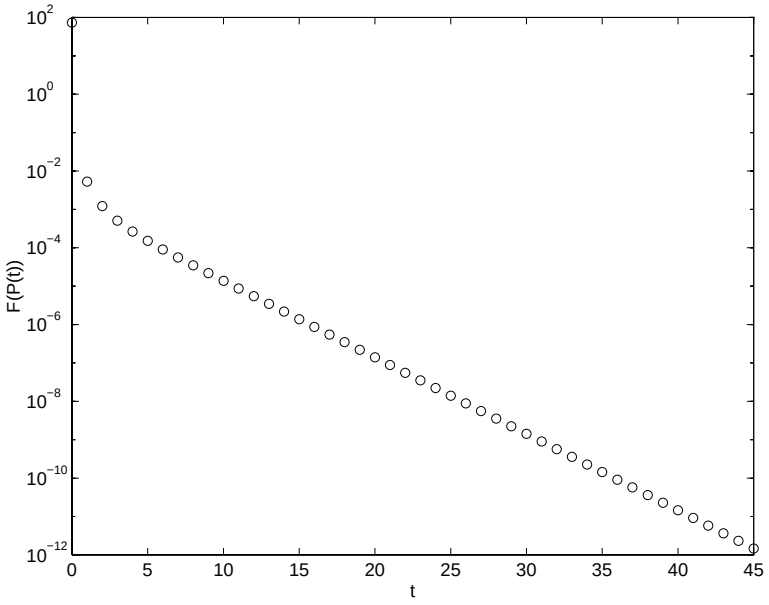


FIG. 3. *History of $F(P(t))$ in Example 1 when $P(t)$ becomes singular.*

We can see also from Figure 3 that this limit point $P(\infty)$ is reducing the objective function (10) to zero. This limit point would be a global minimizer were it not becoming singular. The significant difference here is that since $P(\infty)$ is singular, the corresponding limit point (X, Y) is no longer congruently equivalent to (A, B). In particular, Y is now only positive semidefinite and hence the information of generalized eigenvalues is lost.

Results like this might be disappointing but are still of some theoretic value. It illustrates how congruence transformation in reducing the off-diagonal elements of matrices can go wrong. Our method may be far from practical per se among the many other ways to solve the generalized eigenvalue problem. But readers are reminded that the above illustration of solving Problem 2 by (13) is just one application of our general approach.

It is worthy to remark on three possible remedies along our notion above:

1. The QZ flow [3] is another differential equation approach that is analogous to the steepest descent flow described in this paper. The QZ flow, using orthogonal equivalence transformations instead, does not suffer from the fault of becoming singularity. The symmetric definiteness, however, is not maintained.
2. Even with the descent flow approach, the singularity could be avoided by changing the initial value $P(0)$ and hence taking another path (and there are, indeed, infinitely many such initial guesses.) One could also use flows defined by (23) or (25) to carry out the computation, but we hasten to point out that because either $P(t)P(t)^T$ or $P(t)^T P(t)$ is constant for all t in these cases, not all symmetric-definite pairs (A, B) can be simultaneously diagonalized in this way.

3. Finally, it is possible to avoid the singularity by imposing penalties for singularity in the objective function (10) like those in [1, 9, 19, 20] to avoid the semidefiniteness. This approach will eventually lead to the so-called interior-point methods that have been studied and developed extensively.

Example 2. In general, an inverse eigenvalue problem like Problem 3 can hardly have an exact solution at all. So an approximate solution in the sense of least squares is sometimes desirable. In this case the globally convergent flow defined by (30) becomes particularly meaningful. The flow approach guarantees convergence to a local solution.

To illustrate how the dynamical system (30) behaves, we first generate test data by considering a randomly generated symmetric-definite pair $(\hat{A}, \tilde{B})$:

$$\hat{A} = \begin{bmatrix} -2.8645 & 1.8576 & -2.1532 & 0.6710 & 0.5092 \\ 1.8576 & -0.1855 & 0.5149 & 2.1096 & -1.3318 \\ -2.1532 & 0.5149 & 1.3880 & -0.4591 & 0.3603 \\ 0.6710 & 2.1096 & -0.4591 & -4.3183 & -1.2334 \\ 0.5092 & -1.3318 & 0.3603 & -1.2334 & -1.8954 \end{bmatrix},$$

$$\tilde{B} = \begin{bmatrix} 6.0810 & -2.6691 & 0.6390 & -0.5509 & -1.0124 \\ -2.6691 & 5.5185 & 1.1005 & 0.8248 & 0.8014 \\ 0.6390 & 1.1005 & 2.4625 & 1.9543 & -0.4839 \\ -0.5509 & 0.8248 & 1.9543 & 4.2586 & -0.0535 \\ -1.0124 & 0.8014 & -0.4839 & -0.0535 & 0.8230 \end{bmatrix}.$$

We use its generalized eigenvalues

$$\sigma(\hat{A}, \tilde{B}) = \{3.9955, 0.3093, -0.6662, -1.2920, -3.2878\}$$

as the target spectrum in our experiment. We use $\tilde{A} = \hat{A} - \mathrm{diag}(\hat{A})$ and $\tilde{B}$ as the test data for Problem 3. Apparently, $\mathrm{diag}(\hat{A})$ is one global solution.

Using differential system (30) with initial value

$$P(0) = \begin{bmatrix} 2.4660 & -1.0824 & 0.2591 & -0.2234 & -0.4106 \\ 0 & 2.0849 & 0.6624 & 0.2796 & 0.1713 \\ 0 & 0 & 1.3988 & 1.3061 & -0.3510 \\ 0 & 0 & 0 & 1.5571 & 0.1704 \\ 0 & 0 & 0 & 0 & 0.6877 \end{bmatrix}$$

which comes from the Cholesky decomposition of $\tilde{B}$ (see (29)), we calculate the flow $P(t)$. At convergence we convert $P(\infty)$ into $X(\infty)$ and obtain

$$X(\infty) \approx \begin{bmatrix} 7.1728 & 1.8576 & -2.1532 & 0.6710 & 0.5092 \\ 1.8576 & -0.0080 & 0.5149 & 2.1096 & -1.3318 \\ -2.1532 & 0.5149 & -0.9992 & -0.4591 & 0.3602 \\ 0.6710 & 2.1096 & -0.4591 & -3.9520 & -1.2334 \\ 0.5092 & -1.3318 & 0.3603 & -1.2334 & -2.0060 \end{bmatrix}.$$

We note that the off-diagonal elements of $X(\infty)$ agree with those of $\tilde{A}$ up to the integration error. Therefore, the local solution $\mathrm{diag}(X(\infty))$ we have found is also a global solution. It is interesting to note that $\mathrm{diag}(X(\infty)) \neq \mathrm{diag}(\hat{A})$, indicating that Problem 3 may have multiple solutions. The history of convergence is in Figure 4.

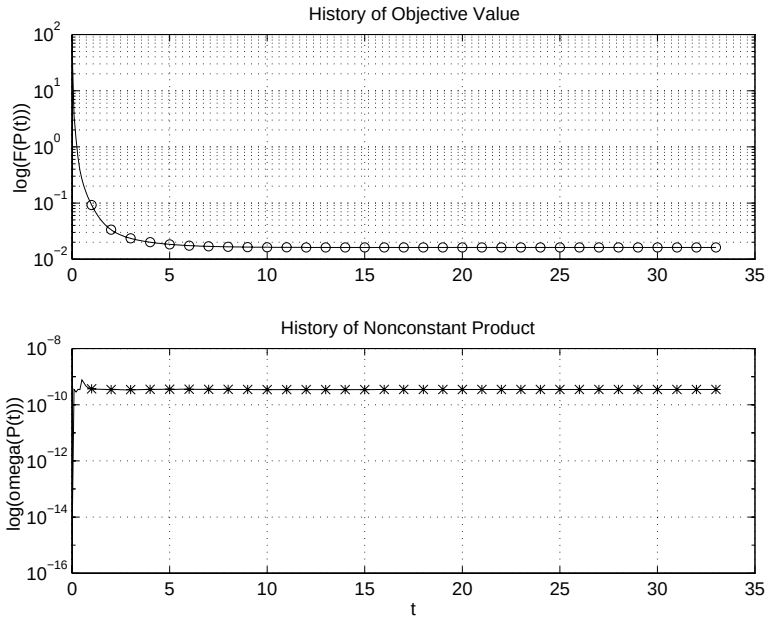


FIG. 4. *History of $F(P(t))$ in Example 2 reaching a global solution.*

Theoretically, it should be that $P(t)^T P(t) = \tilde{B}$ for all t. Numerical calculation introduces errors. For this reason, we closely watch for the values of

$$\omega(P(t)) := \|P(t)^T P(t) - \tilde{B}\|. \tag{31}$$

The second graph in Figure 4 indicates that the discrepancy between theoretical expectation and numerical computation is within our tolerance.

Example 3. We want to stress that the optimization problem (5) is nonlinear and nonconvex. Generally, we cannot expect from any method the luck of hitting the *global* minimizer of any nonlinear or nonconvex optimization problem by one random starting point. One nice feature of our approach, however, is that we are guaranteed to find a local minimizer regardless of where we start and we have plenty of choices of starting points. While it would be nicer to be able to foretell which point/region would serve better as a starting value than the other, the success of such an exploration is perhaps too much to expect due to the nonlinear and nonconvex nature of the problem. On the other hand, since we literally can start from anywhere (e.g., any orthogonal matrix in (29)), we find it is possible, though not the best way, to fish for a "better" starting point by trial and error. We obtain the following results from such a procedure. We have performed many other tests (for the case where a global solution is known to exist) and are always able to find the appropriate starting points after several trials. We have written our code with the convenience of repeated experiments in mind and will make it available upon request.

We report below a case that we think is more challenging than most of the other cases we have tested. Suppose we repeat the experiment in Example 2 with the test

data

$$\tilde{A} = \begin{bmatrix} 1.4637 & -0.3440 & 0.6314 & 0.3603 & 1.2990 \\ -0.3440 & -4.1759 & -0.0370 & 0.8424 & -2.5164 \\ 0.6314 & -0.0370 & -0.5261 & 3.1094 & -0.2112 \\ 0.3603 & 0.8424 & 3.1094 & 2.6428 & -0.9722 \\ 1.2990 & -2.5164 & -0.2112 & -0.9722 & -2.0921 \end{bmatrix},$$

$$\tilde{B} = \begin{bmatrix} 2.1437 & 1.7880 & -0.1595 & 0.7567 & -0.0391 \\ 1.7880 & 7.3264 & -2.8274 & -0.0856 & -0.0528 \\ -0.1595 & -2.8274 & 3.8262 & -1.8245 & -1.7653 \\ 0.7567 & -0.0856 & -1.8245 & 5.0857 & 0.4600 \\ -0.0391 & -0.0528 & -1.7653 & 0.4600 & 1.5725 \end{bmatrix},$$

and the target eigenvalues $\sigma(\tilde{A}, \tilde{B}) = \{2.4562, 1.3627, -0.2342, -0.4489, -250.9816\}$. This time the ratio of the eigenvalues of the largest magnitude to the smallest is relatively large and we expect difficulty.

Suppose we start with the upper triangular matrix in the Cholesky decomposition of $\tilde{B}$, i.e., suppose we choose $U = I$ in (29). At convergence we obtain

$$X(\infty) \approx \begin{bmatrix} 2.9383 & -0.3450 & 0.6401 & 0.3600 & 1.2989 \\ -0.3450 & -13.6834 & -0.0605 & 0.8413 & -2.5144 \\ 0.6401 & -0.0605 & -0.2814 & 2.9862 & -0.2303 \\ 0.3600 & 0.8413 & 2.9862 & 0.3243 & -0.9734 \\ 1.2989 & -2.5144 & -0.2303 & -0.9734 & 0.1012 \end{bmatrix}.$$

Note that the off-diagonal elements of $X(\infty)$ are close, but not within the expected integration error, to those of $\tilde{A}$. From Figure 5 we are convinced that we have reached only a local solution, although that solution is quite close to a global solution. We have checked that $\sigma(X(\infty), \tilde{B})$ agrees with $\sigma(\tilde{A}, \tilde{B})$ up to the integration error.

This example illustrate another difficulty associated with Problem 3. We know that in Problem 3 only the diagonal elements of $\tilde{A}$ are allowed to vary. The off-diagonal elements of $\tilde{A}$ are not supposed to change, but we find that is not the case in our $X(\infty)$. Suppose we project $X(\infty)$ down to the affine subspace of $\tilde{A}$ plus all diagonal matrices to maintain the off-diagonal elements. The eigenvalues of the corresponding projected pair are given by

$$\sigma(\text{off-diag}(\tilde{A}) + \text{diag}(X(\infty)), \tilde{B}) = \{2.4535, 1.4392, -0.2210, -0.4673, -245.6114\}.$$

These values again are close but not within the integration error to the desired target eigenvalues. In other words, this example demonstrates a case where the spectral constraint and the structural constraint cannot be satisfied simultaneously by a local solution.

Suppose we change the starting value to

$$P(0) = \begin{bmatrix} -0.4186 & 0.4414 & -0.7581 & 1.3847 & -0.1868 \\ 0.3510 & 0.1044 & 0.8450 & -0.8140 & -0.5692 \\ -0.6032 & -2.4090 & 0.3297 & 0.3488 & 0.4427 \\ -1.1340 & -0.7609 & 0.9793 & -1.2783 & -0.6188 \\ 0.4421 & -0.8593 & 1.2123 & 0.8660 & -0.7967 \end{bmatrix}$$

which is obtained by multiplying a specific orthogonal matrix (acquired by random trials) to the upper triangular matrix in the Cholesky decomposition of $\tilde{B}$ (see (29)).

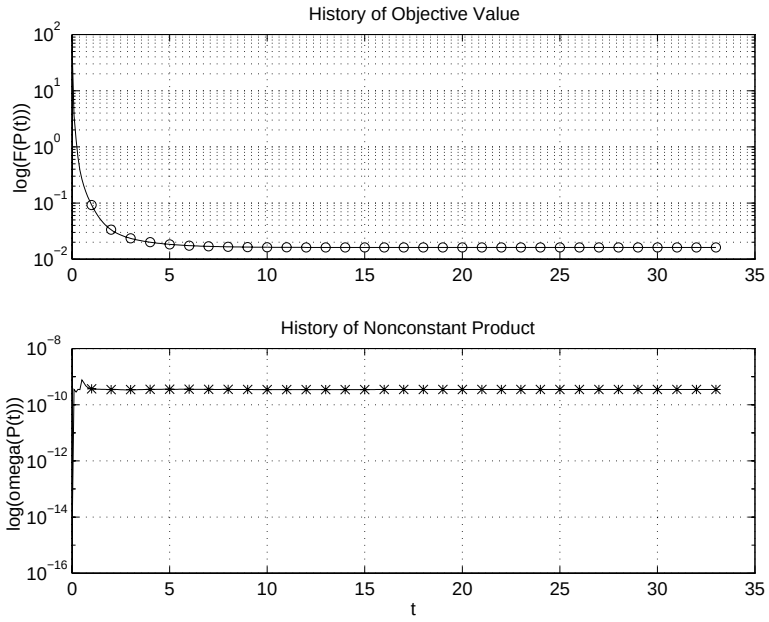


FIG. 5. *History of $F(P(t))$ in Example* 3 *reaching a local solution.*

It turns out that we are able to find a global solution

$$X(\infty) = \begin{bmatrix} 1.4238 & -0.3440 & 0.6315 & 0.3603 & 1.2990 \\ -0.3440 & -4.1785 & -0.0370 & 0.8424 & -2.5164 \\ 0.6315 & -0.0370 & -0.8734 & 3.1093 & -0.2113 \\ 0.3603 & 0.8424 & 3.1093 & 3.0295 & -0.9722 \\ 1.2990 & -2.5164 & -0.2113 & -0.9722 & -1.8037 \end{bmatrix}$$

that satisfies both the spectral and the structural constraints. The history of integration is plotted in Figure 6. The much longer length of integration required for convergence is perhaps due to the stiffness.

6. Conclusion. We have proposed a general framework for the least squares approximation of symmetric-definite pencils subject to generalized eigenvalue constraints. We have illustrated how this approach can be adapted to different applications, including the inverse generalized eigenvalue problems. Although Problem 2 has already enjoyed efficient and reliable numerical algorithms, there are few methods available for Problem 1 and Problem 3. Our approach unifies these different problems under the same framework. The versatility of our method by specifying V_1 and V_2 seem quite interesting.

We have experimented with several descent flows proposed in this paper by using available ordinary differential equation solvers. Our methods guarantee the global convergence to a local solution. By changing integral paths, a global solution sometimes can be reached. It remains to be studied whether a special-purpose integrator/implementation can be developed to make our approach more efficient.

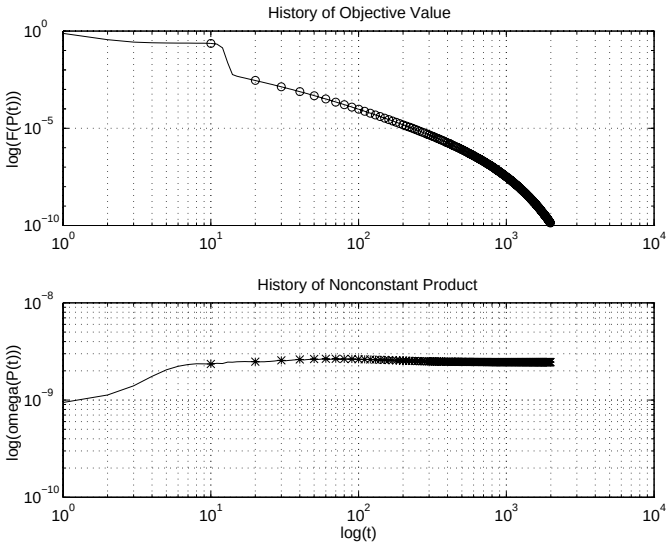


FIG. 6. *History of $F(P(t))$ in Example* 3 *reaching a global solution.*

REFERENCES

[1] S. BOYD, L. E. GHAOUI, E. FERON, AND V. BALAKRISHNAN, *Linear Matrix Inequalities in System and Control Theory*, Stud. Appl. Math., SIAM, Philadelphia, PA, 1994.

[2] X. CHEN AND M. T. CHU, *On the least squares solution of inverse eigenvalue problems*, SIAM J. Numer. Anal., 33 (1996), pp. 2417–2430.

[3] M. T. CHU, *A continuous approximation to the generalized Schur decomposition*, Linear Algebra Appl., 78 (1986), pp. 119–132.

[4] M. T. CHU AND K. R. DRIESSEL, *The projected gradient method for least squares matrix approximations with spectral constraints*, SIAM J. Numer. Anal., 27 (1990), pp. 1050–1060.

[5] M. T. CHU, *A continuous Jacobi-like approach to the simultaneous reduction of real matrices*, Linear Algebra Appl., 147 (1991), pp. 75–96.

[6] J. W. DEMMEL AND A. EDELMAN, *The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms*, Linear Algebra Appl., 230 (1995), pp. 47–60.

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

[8] S. FRIEDLAND, J. NOCEDAL, AND M. L. OVERTON, *The formulation and analysis of numerical methods for inverse eigenvalue problems*, SIAM J. Numer. Anal., 24 (1987), pp. 634–667.

[9] C. C. GONZAGA, *Path-following methods for linear programming*, SIAM Rev., 34 (1992), pp. 167–224.

[10] A. J. HOFFMAN AND H. WIELANDT, *The variation of the spectrum of a normal matrix*, Duke Math. J., 20 (1953), pp. 37–39.

[11] N. G. LLOYD, *Degree Theory*, Cambridge University Press, New York, 1978.

[12] J. MILNOR, *Singular Points of Complex Hyper Surfaces*, Ann. of Math. Stud. 61, Princeton University Press, Princeton, NJ, 1968.

[13] Y. NESTEROV AND A. NEMIROVSKY, *Interior-Point Polynomial Methods in Convex Programming*, Stud. Appl. Math. 13, SIAM, Philadelphia, PA, 1994.

[14] J. D. PRYCE, *Numerical Solution of Sturm-Liouville Problems*, Oxford University Press, New York, 1993.

[15] L. F. SHAMPINE AND M. K. GORDON, *Computer Solution of Ordinary Differential Equations: The Initial Value Problem*, W. H. Freeman, San Francisco, CA, 1975.

[16] L. F. SHAMPINE AND M. W. REICHELT, *The MATLAB ODE Suite*, SIAM J. Sci. Comput., 18 (1997), pp. 1–22.

[17] G. W. STEWART AND J. G. SUN, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.

[18] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.
[19] M. H. Wright, *Interior methods for constrained optimization*, in Acta Numerica 1992, A. Iserles, ed., Cambridge University Press, New York, 1992, pp. 341–407.
[20] M. H. Wright, *Some properties of the Hessian of the logarithmic barrier function*, Math. Programming, 67 (1994), pp. 265–295.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 21–38, January 1998

002

REGULARIZATION OF SINGULAR SYSTEMS BY DERIVATIVE AND PROPORTIONAL OUTPUT FEEDBACK*

D. L. CHU†, H. C. CHAN‡, AND D. W. C. HO‡

Abstract. The problem of the regularization of singular systems by derivative and proportional output feedback is studied. Necessary and sufficient conditions are given to guarantee the existence of a derivative and proportional output feedback such that the closed-loop system is regular and of index at most 1. It is also shown that the closed-loop system becomes strongly controllable and observable by using this feedback.

Key words. regularization, singular systems, output feedback, controllability

AMS subject classifications. 93B05, 93B07, 93B10, 93B40, 93B52, 93C95

PII. S0895479895270963

1. Introduction.

Consider a linear and time-invariant system

$$E\dot{x}(t) = Ax(t) + Bu(t), \qquad y(t) = Cx(t), \tag{1}$$

where $E, A \in \mathbb{R}^{n\times n},\ B \in \mathbb{R}^{n\times m},\ C \in \mathbb{R}^{p\times n}, m \le n,\ p \le n$. When $E = I$, (1) is simply a normal system. Well-known results for normal systems have been obtained over the years and may be found in much literature on control theory. Now our attention will be focused on the case when E is singular.

Existence and uniqueness of solutions to system (1) are guaranteed if (E, A) is regular, that is,

$$\det(\alpha E - \beta A) \not\equiv 0,$$

where the scalars α and β cannot be simultaneously zero. It is well known that for a regular pencil (E, A) there exist nonsingular matrices M and N such that

$$MEN = \begin{bmatrix} I & 0 \\ 0 & J \end{bmatrix}, \qquad MAN = \begin{bmatrix} L & 0 \\ 0 & I \end{bmatrix}, \tag{2}$$

where the eigenvalues of L coincide with the finite eigenvalues of the pencil and J is a nilpotent Jordan matrix such that $J^i = 0,\ \mathrm{J}^{\mathrm{i}-1} \neq 0,\ \mathrm{i} > 0$, corresponding to the infinite eigenvalues. The index of the system, denoted by $\mathrm{ind}(E, A)$, is defined to be equal to the degree i of nilpotency.

For systems that are regular and of index at most 1, they can be separated into purely dynamical and algebraic parts, and in theory the algebraic part can be eliminated to give a reduced-order normal system. The reduction process, however, may be numerically unstable [10].

* Received by the editors February 1, 1995; accepted for publication (in revised form) by P. Van Dooren November 13, 1996. This research was partially supported by the Research Grant Committee (RGC) of Hong Kong project 904065.
http://www.siam.org/journals/simax/19-1/27096.html

† Department of Applied Mathematics, Tsinghua University, Beijing 100084, People's Republic of China.

‡ Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong (madaniel@cityu.edu.hk).

When $\text{ind}(E, A) > 1$, impulses can arise in the response of the system if the control is not sufficiently smooth. Besides, the system can lose causality [15]. Therefore, an appropriate feedback control should be chosen to ensure that the closed-loop system is regular and of index less than or equal to 1.

The eigenstructure of the matrix pencil

$$\alpha E - \beta A \qquad \text{for some} \qquad \alpha, \beta \in \mathbb{R}$$

determines the response of the system. On applying combined derivative and proportional output feedback

$$u = Fy - G\dot{y} + v$$

to (1), the closed-loop system pencil becomes

$$\alpha(E + BGC) - \beta(A + BFC).$$

The main objective of this study is to derive conditions which guarantee the existence of a matrix pair (F, G) such that $(E + BGC, A + BFC)$ is regular and of index at most 1.

For state feedback regularization of singular systems (i.e., $C = I$), numerous studies [5, 6, 7, 11, 12, 13, 16] have been carried out. However, as [1, 2, 8] pointed out, methods described in these papers are based on the Kronecker canonical decomposition of the matrix pencil (E, A), and the system is separated into *fast* and *slow* subsystems in order to obtain the feedback controls. This transformation is well known to be computationally unreliable [14].

Recently, [1, 2, 3] have investigated the regularization of system (1) by state feedback (i.e., $C = I$) and provided numerically stable methods for constructing the feedback gain based on orthogonal matrix decompositions. Because of the differences in nature between state feedback and output feedback, results obtained from [1, 2, 3] cannot directly apply to the case of derivative and proportional output feedback.

In a recent paper [4], a condition has been given for output feedback regularization if the rank of $E + BGC$ is larger than or equal to the rank of E. However, the regularization problem for the complete set of possible ranks of $E + BGC$ has not been characterized so far. This problem will be solved in this paper.

Stabilization of singular systems by derivative and proportional output feedback can be achieved by combining the results on regularization given in this paper and pole assignment technique [9] which, at the same time, preserves regularity. Details are illustrated with an example in section 4.

Next, some notations and definitions are introduced. Within this paper, we denote

$$r_e = \text{rank}(E), \qquad r_a = \text{rank}(A), \qquad r_b = \text{rank}(B), \qquad r_c = \text{rank}(C),$$

$$r_{eb} = \text{rank}\,[E\ B], \qquad r_{ec} = \text{rank}\begin{bmatrix} E \\ C \end{bmatrix}, \qquad r_{ebc} = \text{rank}\begin{bmatrix} E & B \\ C & 0 \end{bmatrix},$$

$$\mathcal{S}_{ebc} = \{r \mid r \text{ is an integer satisfying } r_{eb} + r_{ec} - r_{ebc} \le r \le \min(r_{eb}, r_{ec})\}.$$

It can be seen that $\mathcal{S}_{ebc}$ is the set of integer r satisfying

$$r_{eb} + r_{ec} - r_{ebc} \le r \le \min(r_{eb}, r_{ec}),$$

and it consists exactly of $\min(r_{eb}, r_{ec}) - (r_{eb} + r_{ec} - r_{ebc}) + 1$ integers.

The full column rank matrices $S_E,\ S_{EB},\ S_{EC}$ have their columns span

$$\mathcal{N}(E), \qquad \mathcal{N}\left([E \ \ B]\right), \qquad \mathcal{N}\left(\begin{bmatrix} E \\ C \end{bmatrix}\right),$$

respectively, where $\mathcal{N}(Q)$ is the null space of Q.

The full row rank matrices $T_E,\ T_{EB},\ T_{EC}$ have their columns span

$$\mathcal{N}(E^T), \qquad \mathcal{N}\left([E \ \ B]^T\right), \qquad \mathcal{N}\left(\begin{bmatrix} E \\ C \end{bmatrix}^T\right),$$

respectively.

Concepts of controllability and observability may be extended from state variable systems to singular systems. Few important definitions have been mentioned in [2, 3], but these fundamentals are crucial for the discussions in later sections and are included here for completeness.

DEFINITION 1.1. *Let* (E, A) *be regular. System* (1) *is completely controllable* (*C-controllable*) *if and only if*

C0: $\quad \operatorname{rank}[\alpha E - \beta A \ \ B] = n, \qquad \forall\ (\alpha, \beta) \in \mathbb{R}^2 \backslash \{(0,0)\}.$

DEFINITION 1.2. *Let* (E, A) *be regular*; *then system* (1) *is strongly controllable* (*S-controllable*) *if and only if*

C1: $\quad \operatorname{rank}[\lambda E - A \ \ B] = n, \qquad \forall\ \lambda \in \mathbb{R};$

C2: $\quad \operatorname{rank}[E \ \ AS_E \ \ B] = n.$

Observability conditions can be defined in a similar way as the controllability conditions **C0**, **C1**, and **C2**.

DEFINITION 1.3. *Let* $(E,\ A)$ *be regular. System* (1) *is completely observable* (*C-observable*) *if and only if*

O0: $\quad \operatorname{rank}\begin{bmatrix} \alpha E - \beta A \\ C \end{bmatrix} = n, \qquad \forall\ (\alpha, \beta) \in \mathbb{R}^2 \backslash \{(0,0)\}.$

DEFINITION 1.4. *Let* (E, A) *be regular*; *then system* (1) *is strongly observable* (*S-observable*) *if and only if*

O1: $\quad \operatorname{rank}\begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = n, \qquad \forall\ \lambda \in \mathbb{R};$

O2: $\quad \operatorname{rank}\begin{bmatrix} E \\ T_E^H A \\ C \end{bmatrix} = n.$

This paper is arranged as follows. Section 2 gives some useful preliminaries. Section 3 describes the main results. In particular, three necessary and sufficient conditions for derivative and proportional output feedback regularization problem of singular systems are presented. Results related to the concepts of controllability and observability in singular systems are also discussed. A numerical example is given in section 4. Section 5 makes some concluding remarks.

2. Preliminaries. In this section, some useful results are given.

An easy criterion for regularity may be given by the following theorem, as mentioned similarly in [3].

LEMMA 2.1. *Let* $E, A \in \mathbb{R}^{n \times n}$; *then the pencil* (E, A) *is regular and* $\operatorname{ind}(E, A) \leq 1$ *if and only if*

$$\operatorname{rank}[E \ \ AS_E] = n.$$

Remark. The above lemma serves as a handy tool for determining the regularity of a given matrix pencil.

The following result is a simple extension of Lemma 5 in [3].

LEMMA 2.2. *Let $E \in \mathbb{R}^{n\times q}$ and $B \in \mathbb{R}^{n\times m}$. There exist orthogonal matrices Q, U, and V such that*

$$UEV = \begin{bmatrix} \Sigma_1 & 0 & 0 \\ E_{21} & E_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad UBQ = \begin{bmatrix} 0 & 0 \\ \Sigma_B & 0 \\ 0 & 0 \end{bmatrix},$$

where $E_{22} \in \mathbb{R}^{r_b\times(r_e+r_b-r_{eb})}$ has full column rank and $\Sigma_1 \in \mathbb{R}^{(r_{eb}-r_b)\times(r_{eb}-r_b)}$, $\Sigma_B \in \mathbb{R}^{r_b\times r_b}$ are diagonal positive definite matrices. The partitioning in UEV and UBQ is compatible.

Proof of the above lemma is similar to the one given in [3], which readers can consult. Presented next is a new theorem based on Lemma 2.2.

THEOREM 2.3. *Let $E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$. There exist orthogonal matrices U, V, Q, and W such that*

$$UEV = \begin{bmatrix} \Sigma_1 & 0 & 0 & 0 \\ E_{21} & \Sigma_2 & E_{23} & 0 \\ E_{31} & 0 & E_{33} & 0 \\ E_{41} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad UBQ = \begin{bmatrix} 0 & 0 \\ B_1 & 0 \\ B_2 & 0 \\ B_3 & 0 \\ 0 & 0 \end{bmatrix}, \tag{3}$$

$$WCV = \begin{bmatrix} C_{11} & 0 & \Sigma_C & 0 \\ C_{21} & 0 & 0 & 0 \end{bmatrix},$$

where $\Sigma_1, \Sigma_2, \Sigma_C$ are $(r_{eb}-r_b)\times(r_{eb}-r_b), (r_b+r_{ec}-r_{ebc})\times(r_b+r_{ec}-r_{ebc}), (r_{ebc}-r_{eb})\times(r_{ebc}-r_{eb})$ diagonal positive definite matrices, respectively, E_{33} is an $(r_e+r_{ebc}-r_{eb}-r_{ec})\times(r_{ebc}-r_{eb})$ full row rank matrix, $\begin{bmatrix} B_1^T & B_2^T & B_3^T \end{bmatrix}$ is an $r_b \times r_b$ nonsingular matrix, and $\begin{bmatrix} C_{11}^T & C_{21}^T \end{bmatrix}^T$ is a $p \times (r_{eb}-r_b)$ matrix. The partitioning in UBQ, WCV, and UEV is compatible. Moreover,

$$\begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} = U^*\Sigma_B,$$

where U^* and Σ_B are orthogonal and diagonal positive definite matrices, respectively.

Proof. From Lemma 2.2, we know that there exist orthogonal matrices $\tilde{U}, \tilde{V}$, and Q such that

$$\tilde{U}E\tilde{V} = \begin{bmatrix} \Sigma_1 & 0 & 0 \\ \tilde{E}_{21} & \tilde{E}_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \tilde{U}BQ = \begin{bmatrix} 0 & 0 \\ \Sigma_B & 0 \\ 0 & 0 \end{bmatrix}, \tag{4}$$

where $\tilde{E}_{22} \in \mathbb{R}^{r_b\times(r_e+r_b-r_{eb})}$ has full column rank. Partition

$$\tilde{V} = \begin{bmatrix} \tilde{V}_1 & \tilde{V}_2 \end{bmatrix},$$

where $\tilde{V_1} \in \mathbb{R}^{n\times(r_{eb}-r_b)}$. Let

$$\hat{E} = \begin{bmatrix} \tilde{E}_{22} & 0 \end{bmatrix} \qquad \text{and} \qquad \hat{B} = C\tilde{V_2},$$

where $\hat{E} \in \mathbb{R}^{r_b\times(n-r_{eb}+r_b)}$ and $\hat{B} \in \mathbb{R}^{p\times(n-r_{eb}+r_b)}$. Applying Lemma 2.2 to $\hat{E}$ and $\hat{B}$ once more, we obtain orthogonal matrices U^*, V^*, and W such that

$$U^*\hat{E}V^* = \begin{bmatrix} \Sigma_2 & E_{23} & 0 \\ 0 & E_{33} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad W\hat{B}V^* = \begin{bmatrix} 0 & \Sigma_C & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{5}$$

If we let

$$y_1 = \text{rank}\begin{bmatrix} \hat{E} \\ \hat{B} \end{bmatrix} - \text{rank}\hat{B}, \quad z_c = \text{rank}\hat{B}, \quad y_2 = \text{rank}\hat{E} - y_1,$$

then $\Sigma_2 \in \mathbb{R}^{y_1\times y_1}$ and $\Sigma_C \in \mathbb{R}^{z_c\times z_c}$ are diagonal positive definite matrices, and $E_{33} \in \mathbb{R}^{y_2\times z_c}$ has full row rank. Hence we have the orthogonal matrices

$$U = \begin{bmatrix} I & 0 & 0 \\ 0 & U^* & 0 \\ 0 & 0 & I \end{bmatrix}\tilde{U}, \qquad V = \tilde{V}\begin{bmatrix} I & 0 \\ 0 & V^* \end{bmatrix},$$

and Q, W which give the desired transformation (3). Since

$$r_{ebc} = \text{rank}\begin{bmatrix} E & B \\ C & 0 \end{bmatrix} = \text{rank}\begin{bmatrix} \Sigma_1 & 0 & 0 & 0 \\ \tilde{E}_{21} & \hat{E} & \Sigma_B & 0 \\ 0 & 0 & 0 & 0 \\ C\tilde{V}_1 & 0 & \hat{B} & 0 \end{bmatrix},$$

then

$$z_c = r_{ebc} - r_{eb}, \qquad y_1 = r_{ec} - \text{rank }\Sigma_1 - z_c = r_b + r_{ec} - r_{ebc},$$

$$y_2 = \text{rank }\hat{E} - y_1 = r_e + r_{ebc} - r_{eb} - r_{ec}.$$

This completes the proof. □

The next theorem characterizes the complete set of possible ranks of $E + BGC$.

THEOREM 2.4. *Let $E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, for any integer r satisfying*

$$r_{eb} + r_{ec} - r_{ebc} \;\le\; r \;\le\; \min(r_{eb}, r_{ec});$$

there exists $G_0 \in \mathbb{R}^{m\times p}$ such that

$$\text{rank }(E + BG_0C) = r.$$

Or, equivalently,

$$\{\text{rank}(E + BGC) | G \in \mathbb{R}^{m\times p}\} = \mathcal{S}_{ebc}.$$

Proof. From Theorem 2.3, there exist orthogonal matrices U, V, Q, and W such that E, B, and C are transformed like (3). For any $G \in \mathbb{R}^{m\times p}$, let

$$\tilde{G} = Q^T G W^T = \begin{bmatrix} \tilde{G}_1 & \tilde{G}_2 \\ \tilde{G}_3 & \tilde{G}_4 \end{bmatrix};$$

a direct calculation yields

$$\begin{aligned} \operatorname{rank}(E + BGC) &= \operatorname{rank}\,\Sigma_1 + \operatorname{rank}\,\Sigma_2 + \operatorname{rank}\begin{bmatrix} E_{33} + B_2\tilde{G}_1\Sigma_C \\ B_3\tilde{G}_1\Sigma_C \end{bmatrix} \\ &= r_{eb} + r_{ec} - r_{ebc} + \operatorname{rank}\begin{bmatrix} E_{33} + B_2\tilde{G}_1\Sigma_C \\ B_3\tilde{G}_1\Sigma_C \end{bmatrix}. \end{aligned} \tag{6}$$

Moreover, we have

$$\begin{bmatrix} E_{33} + B_2\tilde{G}_1\Sigma_C \\ B_3\tilde{G}_1\Sigma_C \end{bmatrix} = \hat{A} + \begin{bmatrix} B_2 \\ B_3 \end{bmatrix}\tilde{G}_1\Sigma_C, \tag{7}$$

where

$$\hat{A} = \begin{bmatrix} E_{33} \\ 0 \end{bmatrix}.$$

We can choose

$$\tilde{G}_1 = \Sigma_B^{-1}(U^*)^T\left(\begin{bmatrix} 0 \\ X \end{bmatrix} - \begin{bmatrix} 0 \\ \hat{A} \end{bmatrix}\right)\Sigma_C^{-1}, \tag{8}$$

where $X \in \mathbb{R}^{(r_{ebc}-r_{ec})\times(r_{ebc}-r_{eb})}$ is any matrix satisfying

$$0 \;\le\; i = \operatorname{rank}\,X \;\le\; \min(r_{ebc} - r_{ec},\; r_{ebc} - r_{eb}).$$

Substituting (8) into (7), we obtain

$$\begin{aligned} \operatorname{rank}\begin{bmatrix} E_{33} + B_2\tilde{G}_1\Sigma_C \\ B_3\tilde{G}_1\Sigma_C \end{bmatrix} &= \operatorname{rank}\left(\hat{A} + \begin{bmatrix} B_2 \\ B_3 \end{bmatrix}\Sigma_B^{-1}(U^*)^T\left(\begin{bmatrix} 0 \\ X \end{bmatrix} - \begin{bmatrix} 0 \\ \hat{A} \end{bmatrix}\right)\Sigma_C^{-1}\Sigma_C\right) \\ &= \operatorname{rank}\,X = i. \end{aligned}$$

Therefore

$$0 \;\le\; \operatorname{rank}\begin{bmatrix} E_{33} + B_2\tilde{G}_1\Sigma_C \\ B_3\tilde{G}_1\Sigma_C \end{bmatrix} \;\le\; \min(r_{ebc} - r_{ec},\; r_{ebc} - r_{eb}). \tag{9}$$

Adding $r_{eb} + r_{ec} - r_{ebc}$ to the whole inequality (9), we have the required bound

$$r_{eb} + r_{ec} - r_{ebc} \;\le\; r \;\le\; \min(r_{eb}, r_{ec}),$$

where r is the rank of $E + BGC$. ∎

Remark. Note that the full derivative output feedback matrix is

$$G = Q\tilde{G}W = Q\begin{bmatrix} \tilde{G}_1 & \tilde{G}_2 \\ \tilde{G}_3 & \tilde{G}_4 \end{bmatrix}W,$$

where $\tilde{G}_1$ is given by (8) and $\tilde{G}_2$, $\tilde{G}_3$, and $\tilde{G}_4$ are arbitrarily chosen.

Let

$$\hat{V} = V \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ -\Sigma_C^{-1} C_{11} & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}, \qquad \tilde{E}_{j1} = E_{j1} - E_{j3}\Sigma_C^{-1} C_{11}, \qquad j = 2, 3. \tag{10}$$

Then

$$UE\hat{V} = \begin{bmatrix} \Sigma_1 & 0 & 0 & 0 \\ \tilde{E}_{21} & \Sigma_2 & E_{23} & 0 \\ \tilde{E}_{31} & 0 & E_{33} & 0 \\ E_{41} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad WC\hat{V} = \begin{bmatrix} 0 & 0 & \Sigma_C & 0 \\ C_{21} & 0 & 0 & 0 \end{bmatrix}. \tag{11}$$

Also let

$$UA\hat{V} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \\ A_{51} & A_{52} & A_{53} & A_{54} \end{bmatrix}, \tag{12}$$

where $\hat{V}$ is given in (10). Partitioning of $UE\hat{V}$ and $UA\hat{V}$ is compatible.

THEOREM 2.5. *If*

$$\operatorname{rank} [E \;\; AS_{EC} \;\; B] = \operatorname{rank} \begin{bmatrix} E & AS_{EC} \\ C & 0 \end{bmatrix} = n,$$

then

$$A_{54} \text{ and } \begin{bmatrix} \Sigma_1 & A_{14} \\ \tilde{E}_{31} & A_{34} \\ E_{41} & A_{44} \\ 0 & A_{54} \\ C_{21} & 0 \end{bmatrix}$$

have full row rank and full column rank, respectively.

Proof. If $UE\hat{V}$ and $UA\hat{V}$ are defined by (11) and (12), respectively, then

$$\operatorname{rank} [E \;\; AS_{EC} \;\; B] = n \implies \operatorname{rank} \begin{bmatrix} \Sigma_1 & 0 & 0 & 0 & A_{14} & 0 & 0 \\ \tilde{E}_{21} & \Sigma_2 & E_{23} & 0 & A_{24} & B_1 & 0 \\ \tilde{E}_{31} & 0 & E_{33} & 0 & A_{34} & B_2 & 0 \\ E_{41} & 0 & 0 & 0 & A_{44} & B_3 & 0 \\ 0 & 0 & 0 & 0 & A_{54} & 0 & 0 \end{bmatrix} = n.$$

Since $\begin{bmatrix} B_1^T & B_2^T & B_3^T \end{bmatrix}^T$ and Σ_1 are nonsingular, we have

$$\operatorname{rank} \Sigma_1 + \operatorname{rank} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} + \operatorname{rank} A_{54} = n \implies \operatorname{rank} A_{54} = n - r_{eb}.$$

Note that $A_{54} \in \mathbb{R}^{(n-r_{eb})\times(n-r_{ec})}$; thus A_{54} has full row rank.

We also have

$$\operatorname{rank}\begin{bmatrix} E & AS_{EC} \\ C & 0 \end{bmatrix} = n \implies \operatorname{rank}\begin{bmatrix} \Sigma_1 & 0 & 0 & 0 & A_{14} \\ \tilde{E}_{21} & \Sigma_2 & E_{23} & 0 & A_{24} \\ \tilde{E}_{31} & 0 & E_{33} & 0 & A_{34} \\ E_{41} & 0 & 0 & 0 & A_{44} \\ 0 & 0 & 0 & 0 & A_{54} \\ 0 & 0 & \Sigma_C & 0 & 0 \\ C_{21} & 0 & 0 & 0 & 0 \end{bmatrix} = n.$$

Hence

$$\operatorname{rank}\begin{bmatrix} \Sigma_1 & A_{14} \\ \tilde{E}_{31} & A_{34} \\ E_{41} & A_{44} \\ 0 & A_{54} \\ C_{21} & 0 \end{bmatrix} = n - \operatorname{rank}\,\Sigma_2 - \operatorname{rank}\,\Sigma_C = n - r_b - r_{ec} + r_{eb}.$$

Thus, the theorem has been proved. ∎

3. Derivative and proportional output feedback. Without loss of generality, we assume that $r_{ec} \le r_{eb}$ in most of the results presented in this section; however, for cases $r_{ec} > r_{eb}$, similar argument is applied to the dual system $\left(E^T,\ A^T,\ C^T,\ B^T\right)$.

The regularization of a singular system by using derivative and proportional output feedback is studied in this section. Three necessary and sufficient conditions are provided. The first one relates to derivative output feedback, the second one to combined derivative and proportional output feedback with complete set of $\operatorname{rank}(E + BGC)$, and the last one to proportional output feedback.

THEOREM 3.1. *Given $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, and $r_{ec} \le r_{eb}$, then there exists matrix $G \in \mathbb{R}^{m\times p}$ such that the pencil $(E + BGC, A)$ is regular,* $\operatorname{ind}(E + BGC, A) \le 1$, *and* $\operatorname{rank}\,(E + BGC) = r_{ec}$ *if and only if*

$$\operatorname{rank}\,[E \quad AS_{EC} \quad B] = \operatorname{rank}\begin{bmatrix} E & AS_{EC} \\ C & 0 \end{bmatrix} = n. \tag{13}$$

Proof. Necessity. Let

$$G = Q\begin{bmatrix} \tilde{G}_1 & \tilde{G}_2 \\ \tilde{G}_3 & \tilde{G}_4 \end{bmatrix} W$$

be such that pencil $(E + BGC, A)$ is regular, $\operatorname{ind}(E + BGC, A) \le 1$, and

$$\operatorname{rank}\,(E + BGC) = r_{ec}. \tag{14}$$

Substituting (14) into (6), we obtain

$$\operatorname{rank}\begin{bmatrix} E_{33} + B_2\tilde{G}_1\Sigma_c \\ B_3\tilde{G}_1\Sigma_c \end{bmatrix} = r_{ebc} - r_{eb}.$$

By observing the structure of $U(E + BGC)\hat{V}$, it can be deduced that

$$S_{E+BGC} = \hat{V}\begin{bmatrix} 0 \\ I_{n-r_{ec}} \end{bmatrix} = V\begin{bmatrix} 0 \\ I_{n-r_{ec}} \end{bmatrix} = S_{EC};$$

hence,

$$AS_{E+BGC} = AS_{EC} = U^T \begin{bmatrix} A_{14} \\ A_{24} \\ A_{34} \\ A_{44} \\ A_{54} \end{bmatrix}. \tag{15}$$

Since $(E + BGC, A)$ is regular and $\text{ind}(E + BGC, A) \leq 1$, then by Lemma 2.1, we have

$$\text{rank}\,[E + BGC \quad AS_{E+BGC}] = n$$

or, equivalently,

$$\text{rank}\,[E + BGC \quad AS_{EC}] = n.$$

Since

$$[E + BGC \quad AS_{EC}] = [E \quad AS_{EC}] + BG\,[C \quad 0]\,,$$

therefore, using Theorem 2.4, we obtain

$$\min\left(\text{rank}\,[E \quad AS_{EC} \;\; B]\,,\; \text{rank}\begin{bmatrix} E & AS_{EC} \\ C & 0 \end{bmatrix}\right) \;\geq\; n;$$

however, if

$$\text{rank}\,[E \quad AS_{EC} \;\; B] \leq n \qquad \text{and} \qquad \text{rank}\begin{bmatrix} E & AS_{EC} \\ C & 0 \end{bmatrix} \leq n,$$

then

$$\text{rank}\,[E \quad AS_{EC} \;\; B] = n \qquad \text{and} \qquad \text{rank}\begin{bmatrix} E & AS_{EC} \\ C & 0 \end{bmatrix} = n.$$

Hence the necessary conditions for the existence of G have been proved.

Sufficiency. Since (13) holds, Theorem 2.4 gives that there exists matrix $G \in \mathbb{R}^{m\times p}$ such that

$$\text{rank}\,[E + BGC \quad AS_{EC}] = \text{rank}\,([E \quad AS_{EC}] + BG\,[C \quad 0]) = n.$$

By reversing the proof procedure of the *necessity* part and using Theorem 2.4, we obtain

$$\text{rank}(E + BGC) = r_{ec}$$

and

$$\text{rank}\,[E + BGC \quad AS_{E+BGC}] = \text{rank}\,[E + BGC \quad AS_{EC}] = n.$$

Equivalently, we can say that the pencil $(E + BGC, A)$ is regular and $\text{ind}(E + BGC, A) \leq 1$ by Lemma 2.1. □

Theorem 3.1 gives a necessary and sufficient condition for regularizing (1) by derivative output feedback. This condition is also suitable for the combined derivative and proportional feedback case with a complete set of possible ranks of $E + BGC$.

Denote

$$\mathcal{S}_o = \left\{ G \in \mathbb{R}^{m\times p} | (E + BGC, A + BFC) \text{ is regular and } \operatorname{ind}(E + BGC, A + BFC) \le 1 \text{ for some } F \in \mathbb{R}^{m\times p} \right\}.$$

THEOREM 3.2. *Let* $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, *and* $r_{ec} \le r_{eb}$; *then*

$$\{\operatorname{rank}(E + BGC) | G \in \mathcal{S}_o\} = \mathcal{S}_{ebc} \tag{16}$$

if and only if (13) *is true.*

Proof. Necessity. Since $r_{ec} \in \mathcal{S}_{ebc}$, there exists $G \in \mathcal{S}_o$ such that $(E + BGC, A + BFC)$ is regular, $\operatorname{ind}(E + BGC, A + BFC) \le 1$, and $\operatorname{rank}(E + BGC) = r_{ec}$ for some $F \in R^{m\times p}$. From the proof of Theorem 3.1, we have

$$S_{E+BGC} = S_{EC} = V \begin{bmatrix} 0 \\ I_{n-r_{ec}} \end{bmatrix} = \hat{V} \begin{bmatrix} 0 \\ I_{n-r_{ec}} \end{bmatrix};$$

then we obtain

$$(A + BFC)S_{E+BGC} = (A + BFC)S_{EC} = AS_{EC} = AS_{E+BGC}, \quad \forall F \in \mathbb{R}^{m\times p}.$$

Hence

$$\operatorname{rank}(E + BGC, AS_{E+BGC}) = \operatorname{rank}\left[E + BGC, (A + BFC)S_{E+BGC}\right].$$

From Lemma 2.1, it can be concluded that $(E + BGC, A)$ is regular and of index at most 1 and $\operatorname{rank}(E + BGC) = r_{ec}$. Together with Theorem 3.1, (13) results.

Sufficiency. Assume $r \in \mathcal{S}_{ebc}$ is an arbitrary integer. Now that we have

$$\operatorname{rank}\begin{bmatrix} E & AS_{EC} & B \end{bmatrix} = n, \qquad \operatorname{rank}\begin{bmatrix} E & AS_{EC} \\ C & 0 \end{bmatrix} = n;$$

then by Theorem 2.5, it can be deduced that

$$A_{54} \qquad \text{and} \qquad \begin{bmatrix} \Sigma_1 & A_{14} \\ \tilde{E}_{31} & A_{34} \\ E_{41} & A_{44} \\ 0 & A_{54} \\ C_{21} & 0 \end{bmatrix} \tag{17}$$

have full row rank and full column rank, respectively. Let

$$\begin{bmatrix} E_{23} \\ E_{33} \\ 0 \end{bmatrix} = \begin{bmatrix} E_{23}^1 & E_{23}^2 \\ E_{33}^1 & E_{33}^2 \\ 0 & 0 \end{bmatrix}, \quad \tilde{G}_1 = \begin{bmatrix} \tilde{G}_{11} & \tilde{G}_{12} \end{bmatrix}, \quad \Sigma_C = \begin{bmatrix} \Sigma_C^1 & 0 \\ 0 & \Sigma_C^2 \end{bmatrix},$$

where E_{23}^2 and Σ_C^2 are $(r_b + r_{ec} - r_{ebc}) \times (r_{ec} - r)$ and $(r_{ec} - r) \times (r_{ec} - r)$ matrices, respectively. Since $\begin{bmatrix} B_1^T & B_2^T & B_3^T \end{bmatrix}^T$ and Σ_C are nonsingular, we can choose

$$\tilde{G}_{12} = -\Sigma_B^{-1}(U^*)^T \begin{bmatrix} E_{23}^2 \\ E_{33}^2 \\ 0 \end{bmatrix} (\Sigma_C^2)^{-1}; \tag{18}$$

then

$$\begin{bmatrix} E_{23}^2 \\ E_{33}^2 \\ 0 \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix} \tilde{G}_{12}\Sigma_C^2 = 0.$$

Define

$$\begin{bmatrix} A_{13} \\ A_{23} \\ A_{33} \\ A_{43} \\ A_{53} \end{bmatrix} = \begin{bmatrix} A_{13}^1 & A_{13}^2 \\ A_{23}^1 & A_{23}^2 \\ A_{33}^1 & A_{33}^2 \\ A_{43}^1 & A_{43}^2 \\ A_{53}^1 & A_{53}^2 \end{bmatrix}, \qquad \tilde{S}_{E+BGC} = \hat{V}\begin{bmatrix} 0 \\ I_{n-r} \end{bmatrix},$$

$$\tilde{F} = Q^T F W^T = \begin{bmatrix} \tilde{F}_1 & \tilde{F}_2 \\ \tilde{F}_3 & \tilde{F}_4 \end{bmatrix}, \qquad \tilde{F}_1 = \begin{bmatrix} \tilde{F}_{11} & \tilde{F}_{12} \end{bmatrix};$$

then we have

$$E + BGC = U^T \begin{bmatrix} \Sigma_1 & 0 & 0 & 0 & 0 \\ \tilde{E}_{21} + B_1\tilde{G}_2C_{21} & \Sigma_2 & E_{23}^1 + B_1\tilde{G}_{11}\Sigma_C^1 & 0 & 0 \\ \tilde{E}_{31} + B_2\tilde{G}_2C_{21} & 0 & E_{33}^1 + B_2\tilde{G}_{11}\Sigma_C^1 & 0 & 0 \\ E_{41} + B_3\tilde{G}_2C_{21} & 0 & B_3\tilde{G}_{11}\Sigma_C^1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \hat{V}^{-1} \tag{19}$$

and

$$A+BFC = U^T \begin{bmatrix} A_{11} & A_{12} & A_{13}^1 & A_{13}^2 & A_{14} \\ A_{21} + B_1\tilde{F}_2C_{21} & A_{22} & A_{23}^1 + B_1\tilde{F}_{11}\Sigma_C^1 & A_{23}^2 + B_1\tilde{F}_{12}\Sigma_C^2 & A_{24} \\ A_{31} + B_1\tilde{F}_2C_{21} & A_{32} & A_{33}^1 + B_2\tilde{F}_{11}\Sigma_C^1 & A_{33}^2 + B_2\tilde{F}_{12}\Sigma_C^2 & A_{34} \\ A_{41} + B_3\tilde{F}_2C_{21} & A_{42} & A_{43}^1 + B_3\tilde{F}_{11}\Sigma_C^1 & A_{43}^2 + B_3\tilde{F}_{12}\Sigma_C^2 & A_{44} \\ A_{51} & A_{52} & A_{53}^1 & A_{53}^2 & A_{54} \end{bmatrix} \hat{V}^{-1}.$$

Therefore,

$$\operatorname{rank}\begin{bmatrix} E + BGC & (A + BFC)\tilde{S}_{E+BGC} \end{bmatrix} = \operatorname{rank} Y + \operatorname{rank} \Sigma_2,$$

where

$$\begin{aligned} Y &= \begin{bmatrix} \Sigma_1 & 0 & A_{13}^2 & A_{14} \\ \tilde{E}_{31} + B_2\tilde{G}_2C_{21} & E_{33}^1 + B_2\tilde{G}_{11}\Sigma_C^1 & A_{33}^2 + B_2\tilde{F}_{12}\Sigma_C^2 & A_{34} \\ E_{41} + B_3\tilde{G}_2C_{21} & B_3\tilde{G}_{11}\Sigma_C^1 & A_{43}^2 + B_3\tilde{F}_{12}\Sigma_C^2 & A_{44} \\ 0 & 0 & A_{53}^2 & A_{54} \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_1 & 0 & A_{13}^2 & A_{14} \\ \tilde{E}_{31} & E_{33}^1 & A_{33}^2 & A_{34} \\ E_{41} & 0 & A_{43}^2 & A_{44} \\ 0 & 0 & A_{53}^2 & A_{54} \end{bmatrix} + \begin{bmatrix} 0 \\ B_2 \\ B_3 \\ 0 \end{bmatrix} \begin{bmatrix} \tilde{G}_{11} & \tilde{F}_{12} & \tilde{G}_2 \end{bmatrix} \begin{bmatrix} 0 & \Sigma_C^1 & 0 & 0 \\ 0 & 0 & \Sigma_C^2 & 0 \\ C_{21} & 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

is an $(n + r_{ebc} - r_b - r_{ec}) \times (n + r_{ebc} - r_b - r_{ec})$ matrix. From (17), we have

$$\operatorname{rank}\begin{bmatrix} \Sigma_1 & 0 & A_{13}^2 & A_{14} & 0 \\ \tilde{E}_{31} & E_{33}^1 & A_{33}^2 & A_{34} & B_2 \\ E_{41} & 0 & A_{43}^2 & A_{44} & B_3 \\ 0 & 0 & A_{53}^2 & A_{54} & 0 \end{bmatrix} = n + r_{ebc} - r_b - r_{ec}$$

and

$$\operatorname{rank}\begin{bmatrix} \Sigma_1 & 0 & A_{13}^2 & A_{14} \\ \tilde{E}_{31} & E_{33}^1 & A_{33}^2 & A_{34} \\ E_{41} & 0 & A_{43}^2 & A_{44} \\ 0 & 0 & A_{53}^2 & A_{54} \\ 0 & \Sigma_c^1 & 0 & 0 \\ 0 & 0 & \Sigma_c^2 & 0 \\ C_{21} & 0 & 0 & 0 \end{bmatrix} = \operatorname{rank}\Sigma_C + \operatorname{rank}\begin{bmatrix} \Sigma_1 & A_{14} \\ \tilde{E}_{31} & A_{34} \\ E_{41} & A_{44} \\ 0 & A_{54} \\ C_{21} & 0 \end{bmatrix} = n + r_{ebc} - r_b - r_{ec}.$$

From Theorem 2.4, we know that there exists a real matrix $\begin{bmatrix}\tilde{G}_{11} & \tilde{F}_{12} & \tilde{G}_2\end{bmatrix}$ such that

$$\operatorname{rank} Y = n + r_{ebc} - r_b - r_{ec};$$

furthermore,

$$\operatorname{rank}\begin{bmatrix} E + BGC & (A + BFC)\tilde{S}_{E+BGC} \end{bmatrix} = \operatorname{rank}\Sigma_2 + \operatorname{rank} Y = n.$$

Now we have shown the existence of G satisfying

$$\operatorname{rank}(E + BGC) = r,$$

where $r_{eb} + r_{ec} - r_{ebc} \le r \le r_{ec}$. The above equality and (19) imply $S_{E+BGC} = \tilde{S}_{E+BGC}$; then

$$\operatorname{rank}[E + BGC \quad (A + BFC)S_{E+BGC}] = n.$$

Hence from Lemma 2.1,

$$(E + BGC,\ A + BFC) \text{ is regular and } \operatorname{ind}(E + BGC, A + BFC) \le 1.$$

Hence the theorem has been proved. □

Related to proportional output feedback without the assumption of $r_{ec} \le r_{eb}$, we have Theorem 3.3.

THEOREM 3.3. *Let* $E, A \in \mathbb{R}^{n\times n},\ B \in \mathbb{R}^{n\times m},\ C \in \mathbb{R}^{p\times n},\ m \le n,\ p \le n$; *then there exists* $F \in \mathbb{R}^{m\times p}$ *such that* $(E,\ A+BFC)$ *is regular and* $\operatorname{ind}(E,\ A+BFC) \le 1$ *if and only if*

$$\operatorname{rank}[E \quad AS_E \quad B] = \operatorname{rank}\begin{bmatrix} E & AS_E \\ 0 & CS_E \end{bmatrix} = n. \tag{20}$$

Proof. From Lemma 2.1, the pencil $(E,\ A + BFC)$ is regular and $\operatorname{ind}(E,\ A + BFC) \le 1$ if and only if

$$\operatorname{rank}[E \quad (A + BFC)S_E] = \operatorname{rank}([E\ AS_E] + BF[0\ \mathrm{CS_E}]) = n.$$

From Theorem 2.4, the existence of F in the above equation is equivalent to

$$\min\left(\operatorname{rank}[E \quad AS_E \quad B],\ \operatorname{rank}\begin{bmatrix} E & AS_E \\ 0 & CS_E \end{bmatrix}\right) \ge n \tag{21}$$

and

$$\operatorname{rank}[E \quad AS_E \quad B] + \operatorname{rank}\begin{bmatrix} E & AS_E \\ 0 & CS_E \end{bmatrix} - \operatorname{rank}\begin{bmatrix} E & AS_E & B \\ 0 & CS_E & 0 \end{bmatrix} \le n. \tag{22}$$

And in fact

$$\text{rank}[E \;\; AS_E \;\; B] \leq \; n, \qquad \text{rank} \begin{bmatrix} E & AS_E \\ 0 & CS_E \end{bmatrix} \leq n.$$

Inequalities (21) and (22) lead to (20) and the theorem has been proved. □

Remark. In [4], sufficient conditions (i.e., **C1, O1, C2, O2** in their paper) are given to ensure that there exist feedback matrices F and G such that the closed-loop system is strongly controllable and strongly observable, with an index at most 1 and

$$t_1 \;\leq\; \text{rank}\,(E + BGC) \;\leq\; t_1 + t_2 \qquad \text{for some positive integers } t_1, t_2.$$

In this paper, we have obtained three necessary and sufficient conditions (see Theorems 3.1, 3.2, and 3.3) for the regularization of singular system (1) by derivative output feedback, combined derivative and proportional output feedback, and proportional output feedback, respectively, such that the closed-loop system is regular and has an index at most 1 with

$$r_{eb} + r_{ec} - r_{ebc} \;\leq\; \text{rank}\,(E + BGC) \;\leq\; \min(r_{eb},\; r_{ec}). \tag{23}$$

In fact, these lower and upper bounds are reachable.

We will show that the upper bound $\min(r_{eb}, r_{ec})$ achieved in this paper is greater than the one in [4] (i.e., $t_1 + t_2$). By examining the reduced forms of (E, A, B, C) in [4] and (E, B, C) in our work, we can deduce that

$$r_{eb} = t_1 + t_2 + t_3, \qquad r_{ec} = t_1 + t_2 + t_5, \qquad r_{ebc} = t_1 + 2t_2 + t_3 + t_5,$$

where t_3, t_5 are positive integers defined in [4]. After some manipulations, we obtain

$$\begin{aligned} & t_1 + t_2 = r_{eb} - t_3 \;=\; r_{ec} - t_5 \\ \Longrightarrow \;\; & t_1 + t_2 \leq \min(r_{eb},\; r_{ec}). \end{aligned}$$

Hence, under **C1, O1, C2, O2**, and (13), it is obvious that the upper bound in (23)

$$\min(r_{ec}, r_{eb})$$

is greater than the 1 (i.e., $t_1 + t_2$) shown in [4].

3.1. Controllability and observability of singular systems. Issues concerning output feedback regularization problems relating to C-controllability (and C-observability) and S-controllability (S-observability) of singular systems (1) are discussed next.

Obviously, if system (1) is C-controllable (C-observable), it is S-controllable (S-observable). It is known that by using derivative output feedback, system (1) can be transformed into a normal system which is C-controllable and C-observable if and only if system (1) is C-controllable and C-observable. In case system (1) is not C-controllable and C-observable, it is still possible to use derivative output feedback to modify system (1) such that the closed-loop system is regular, S-controllable (or S-observable), and has index at most 1. This fact is illustrated by the following theorem for derivative output feedback.

THEOREM 3.4. *Given system* (1), $r_{ec} \leq r_{eb}$ *(or* $r_{ec} > r_{eb}$*), there exists derivative output feedback* $u = -G\dot{y} + v$ *such that the closed-loop system is S-controllable* (*or*

S-observable), regular, has index at most 1, *and* $\operatorname{rank}(E+BGC)=r_{ec}$ *if and only if* (E,A,B) (*or* (E,A,C)) *satisfies condition* **C1** (*or* **O1**) *and* (13).

Proof. Since

$$\operatorname{rank}\left[\lambda(E+BGC)-A \quad B\right]=\operatorname{rank}\left[\lambda E-A \quad B\right], \quad \forall\lambda\in\mathbb{R}, \tag{24}$$

condition **C1** is preserved by derivative output feedback. Furthermore, by Lemma 2.1, $(E+BGC,A)$ is regular and $\operatorname{ind}(E+BGC,A)\leq 1$ is equivalent to

$$\begin{aligned}&\operatorname{rank}\left[E+BGC \quad AS_{E+BGC}\right]=n\\ \Longrightarrow\ &\operatorname{rank}\left[E+BGC \quad AS_{E+BGC} \quad B\right]=n,\end{aligned} \tag{25}$$

which is simply **C2** for the closed-loop system. Also, results of Theorem 3.1 give (13). Hence this theorem has been proved. □

The above theorem is extended to the case of proportional feedback without proof.

THEOREM 3.5. *Given system* (1), *there exists proportional output feedback* $u=Fy+v$ *such that the closed-loop system is S-controllable* (*or S-observable*), *regular, and has index at most* 1 *if and only if the triple* (E,A,B) (*or* (E,A,C)) *satisfies condition* **C1** (*or* **O1**) *and* (20).

Given system (1), define

$$\begin{aligned}\tilde{\mathcal{S}}=\big\{&(F,G)|F,G\in\mathbb{R}^{m\times p},\ \text{closed-loop system given by applying derivative and}\\ &\text{proportional output feedback } u=Fy-G\dot{y}+v \text{ is S-controllable, regular, and}\\ &\text{has index at most } 1\big\},\\ \tilde{\mathcal{S}}_o=\Big\{&G|(F,G)\in\tilde{\mathcal{S}} \text{ for some } F\in\mathbb{R}^{m\times p}\Big\}.\end{aligned}$$

Then by applying combined derivative and proportional output feedback to (1), similar results can be obtained without proof.

THEOREM 3.6. *Given system* (1), $r_{ec}\leq r_{eb}$, *then*

$$\Big\{\operatorname{rank}(E+BGC)|G\in\tilde{\mathcal{S}}_o\Big\}=\mathcal{S}_{ebc}$$

if and only if the triple (E,A,B) *satisfies condition* **C1** *and* (13).

Remark. For Theorem 3.6 we have a similar result for $r_{eb}<r_{ec}$ with the dual system.

4. An example. In this section, a numerical example is presented to illustrate the results given in the previous section. It should be noted that our main results are derived based on the condensed form given in Theorem 2.3, which can be computed in a numerically stable way. A numerical procedure based on this condensed form is included in Appendix A. The program is coded in MATLAB[1] and performed on a SPARC-10 Sun[2] workstation running under Unix[3].

This test problem is adopted from [5] with slight variations. Recall that in this

[1] MATLAB is a trademark of the Mathworks, Inc.
[2] Sun is a trademark of Sun Microsystems.
[3] Unix is a trademark of AT&T.

example $n = 6,\ m = 2,\ p = 2$, and the system matrices are given by

$$E = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

It can be shown that (E, A) is not regular and of index greater than 1. Besides, $\operatorname{rank}(E) = 3$, that is, the maximum number of open-loop finite poles is 3. The input and output matrices are defined to be

$$B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & -1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Since the system satisfies (13) and

$$r_{eb} = \operatorname{rank}[E \quad B] = 5 = \operatorname{rank}\begin{bmatrix} E \\ C \end{bmatrix} = r_{ec},$$

the possible number of closed-loop finite poles is

$$3 \ \leq \ r \ \leq \ 5.$$

For instance, if we choose $r = 3$, we obtain

$$F = \begin{bmatrix} -0.1060 & -0.6490 \\ -1.0223 & -0.7382 \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

By Lemma 2.1, we can verify that the closed-loop system $(E + BGC, A + BFC)$ now becomes regular and of index at most 1 and $\operatorname{rank}(E + BGC) = 3$. Besides, the system satisfies (20), and Theorem 3.1 is verified.

Similarly for $r = 4$, we obtain

$$F = \begin{bmatrix} 0 & -0.3765 \\ 0 & -0.8450 \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} -0.2614 & 0 \\ -1.0985 & 0 \end{bmatrix}.$$

By Lemma 2.1, we can verify that the closed-loop system $(E + BGC, A + BFC)$ now becomes regular and of index at most 1. Besides, the rank of $E + BGC$ has been increased to 4, which is the number of finite poles for the closed-loop system.

It can be verified that the resulting closed-loop system $(E + BGC, A + BFC)$ possesses unstable eigenvalues. By following the approach in [9], we can find a proportional feedback matrix

$$F_s = \begin{bmatrix} 14.1412 & 0 \\ 0 & 0 \end{bmatrix}$$

such that $(E + BGC, A + BFC + BF_sC)$ is stable. Note that regularity is preserved.

Similarly, for $r = 5$ we obtain

$$F = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad G = \begin{bmatrix} -0.1570 & -0.5034 \\ -0.5115 & -0.7280 \end{bmatrix}.$$

By Lemma 2.1, we can verify that the closed-loop system $(E + BGC, A + BFC)$ now becomes regular and of index at most 1. Besides, the rank of $E + BGC$ has been increased to 5, which is the number of finite poles for the closed-loop system.

It can be verified that the resulting closed-loop system $(E + BGC, A + BFC)$ possesses unstable eigenvalues. By following the approach in [9], we can find a proportional feedback matrix

$$F_s = \begin{bmatrix} -0.3952 & 0.4327 \\ -0.4980 & -0.8184 \end{bmatrix}$$

such that $(E + BGC, A + BFC + BF_sC)$ is stable. Note that regularity is preserved.

Remark. In this paper, we have discussed the issue of finding the pair (F, G) such that $(E+BGC, A+BFC)$ is regular, has index at most 1, and possesses desired number of finite poles. But from the numerical point of view, an *optimal* G is expected such that $E + BGC$ is well conditioned. In the state feedback case, [3] has given a method to solve it. However, in the case of output feedback, this is still an open question and requires further investigation.

5. Conclusions. In this paper, we have studied the problem of the regularization of singular systems by derivative and proportional output feedback. Some necessary and sufficient conditions are given to guarantee the existence of a derivative and proportional output feedback such that the closed-loop system is regular and of index at most 1. It is also shown that the closed-loop system becomes strongly controllable and observable by using this feedback. A numerical example is given to illustrate the result.

Appendix A. A numerical algorithm is developed to implement the main result given in section 3.

Input: $E,\ A \in \mathbb{R}^{n\times n},\ B \in \mathbb{R}^{n\times m},\ C \in \mathbb{R}^{p\times n}$ and an integer r which must be within the following bound

$$r_{eb} + r_{ec} - r_{ebc} \ \le\ r\ \le \min(r_{eb},\ r_{ec}).$$

Output: $F,\ G \in \mathbb{R}^{m\times p}$ such that

(a) the pencil $(E + BGC,\ A + BFC)$ is regular;
(b) $\text{ind}(E + BGC,\ A + BFC) \le 1$; and
(c) $\text{rank}\ (E + BGC) = r$.

STEP 1. If $r_{ec} \le r_{eb}$, proceed to the next step. Otherwise, use the dual system for further manipulations; that is, let

$$E = E^T,\quad A = A^T,\quad B = C^T,\quad C = B^T.$$

STEP 2. Check if (13) holds.

STEP 3. Find orthogonal matrices $\tilde{U},\ \tilde{V} \in \mathbb{R}^{n\times n},\ Q \in \mathbb{R}^{m\times n}$ such that (4) is satisfied.

STEP 4. Form $\hat{V}$, $UE\hat{V}$, and $UA\hat{V}$ using (10), (11), and (12). Partition the matrices

$$\begin{bmatrix} E_{23} \\ E_{33} \end{bmatrix} = \begin{bmatrix} E_{23}^1 & E_{23}^2 \\ E_{33}^1 & E_{33}^2 \end{bmatrix}, \quad \begin{bmatrix} A_{13} \\ A_{23} \\ A_{33} \\ A_{43} \\ A_{53} \end{bmatrix} = \begin{bmatrix} A_{13}^1 & A_{13}^2 \\ A_{23}^1 & A_{23}^2 \\ A_{33}^1 & A_{33}^2 \\ A_{43}^1 & A_{43}^2 \\ A_{53}^1 & A_{53}^2 \end{bmatrix}, \ \Sigma_C = \begin{bmatrix} \Sigma_C^1 & 0 \\ 0 & \Sigma_C^2 \end{bmatrix},$$

where $E_{23}^2 \in \mathbb{R}^{(r_b+r_{ec}-r_{ebc})\times(r_{ec}-r)}$ and $\Sigma_C^2 \in \mathbb{R}^{(r_{ec}-r)\times(r_{ec}-r)}$. Note that the partitioning is compatible.

STEP 5. Find the matrix $\begin{bmatrix} \tilde{G}_{11} & \tilde{F}_{12} & \tilde{G}_2 \end{bmatrix}$ such that

$$\tilde{B}\begin{bmatrix} \tilde{G}_{11} & \tilde{F}_{12} & \tilde{G}_2 \end{bmatrix}\tilde{C} + \tilde{A}$$

is nonsingular. Here

$$\tilde{B} = \begin{bmatrix} 0 \\ B_2 \\ B_3 \\ 0 \end{bmatrix}, \ \tilde{C} = \begin{bmatrix} 0 & \Sigma_C^1 & 0 & 0 \\ 0 & 0 & \Sigma_C^2 & 0 \\ C_{21} & 0 & 0 & 0 \end{bmatrix}, \ \tilde{A} = \begin{bmatrix} \Sigma_1 & 0 & A_{13}^2 & A_{14} \\ \tilde{E}_{31} & E_{33}^1 & A_{33}^2 & A_{34} \\ E_{41} & 0 & A_{43}^2 & A_{44} \\ 0 & 0 & A_{53}^2 & A_{54} \end{bmatrix}$$

and $\tilde{G}_{11}, \tilde{F}_{12}, \tilde{G}_2$ are $r_b \times (r_{ebc} - r_{eb} - r_{ec} + r), r_b \times (r_{ec} - r), r_b \times (p - r_{ebc} + r_{eb})$ matrices, respectively.

STEP 6. Compute $\tilde{G}_{12}$ according to (18).

STEP 7. The desired output feedback matrices are

$$F = Q\begin{bmatrix} \tilde{F}_1 & \tilde{F}_2 \\ \tilde{F}_3 & \tilde{F}_4 \end{bmatrix}W, \quad G = Q\begin{bmatrix} \tilde{G}_1 & \tilde{G}_2 \\ \tilde{G}_3 & \tilde{G}_4 \end{bmatrix}W,$$

where $\tilde{F}_1 = [\tilde{F}_{11} \ \ \tilde{F}_{12}]$, $\tilde{G} = [\tilde{G}_{11} \ \ \tilde{G}_{12}]$, and $\tilde{F}_{11}, \ \tilde{F}_2, \ \tilde{F}_3, \ \tilde{F}_4, \ \tilde{G}_3, \ \tilde{G}_4$ are arbitrarily chosen.

STEP 8. If $(E+BGC, \ A+BFC)$ is unstable, find F_s by [9] such that $(E+BGC, \ A+BFC+BF_sC)$ is stable and regular. Then set $F = F + F_s$.

Remarks.

- Step 3 can be achieved by using the procedure given in the proof of Lemma 5 in [3]. This procedure only requires some simple matrix manipulations and applying SVD or QR methods for a few times, which makes this procedure numerically reliable.
- Step 5 can be achieved as follows: apply Theorem 2.3 to $(\tilde{A}, \ \tilde{B}, \ \tilde{C})$ to get a condensed form like (3); then use the procedure provided in the proof of Theorem 2.4 to get the desired matrix.
- Steps 4, 5, and 6 only require simple matrix manipulations and the inverses of two diagonal and positive-definite matrices which can be computed in a numerically reliable way.

Acknowledgment. We gratefully acknowledge the anonymous referees and editors for their kind and detailed comments on the early version of this paper. The research reported here also benefitted a great deal from the valuable suggestions given by them.

REFERENCES

[1] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols, *Derivative feedback for descriptor systems*, presented at the IFAC workshop on System Structure and Control: State–Space and Polynomial Methods, Prague, September 1989.

[2] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols, *On derivative and proportional feedback design for descriptor systems*, in Proc. Internat. Symp. MTNS-89, Vol. III, Birkäuser, Basel, 1990, pp. 437–446.

[3] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols, *Regularization of descriptor systems by derivative and proportional state feedback*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 46–67.

[4] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols, *Regularization of descriptor systems by output feedback*, IEEE Trans. Automat. Control, 39 (1994), pp. 1742–1748.

[5] M. A. Christodoulou, *Decoupling in the design and synthesis of singular systems*, Automatica, 22 (1986), pp. 245–249.

[6] L. Dai, *Impulsive modes and causality in singular systems*, Internat. J. Control, 50 (1989), pp. 1267–1281.

[7] L. Dai, *Singular Control Systems*, Lecture Notes in Control and Information Sciences 118, Springer–Verlag, Berlin, 1989.

[8] L. R. Fletcher, J. Kautsky, and N. K. Nichols, *Eigenstructure assignment in descriptor systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 1138–1141.

[9] L. R. Fletcher, *Eigenstructure assignment by output feedback in descriptor systems*, IEEE Proc. Pt. D, 135 (1988), pp. 302–308.

[10] V. L. Mehrmann, *The Autonomous Linear Quadratic Control Problem: Theory and Numerical Solutions*, Lecture Notes in Control and Information Sciences 163, Springer–Verlag, Berlin, 1991.

[11] R. Mukundan and W. Dayawansa, *Feedback control of singular systems–proportional and derivative feedback of the state*, Internat. J. Systems Sci., 14 (1983), pp. 615–632.

[12] K. Özçaldiran and F. L. Lewis, *On the regularizability of singular systems*, IEEE Trans. Automat. Control, 35 (1990), pp. 1156–1160.

[13] M. A. Shayman and Z. Zhou, *Feedback control and classification of generalized linear systems*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 483–494.

[14] P. M. Van Dooren, *The generalized eigenstructure problem in linear system theory*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 111–129.

[15] E. L. Yip and R. F. Sincovec, *Solvability, controllability, and observability of continuous descriptor systems*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 702–707.

[16] Z. Zhou, M. A. Shayman, and T. J. Tarn, *Singular systems: A new approach in the time domain*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 42–50.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 39–65, January 1998

003

PERTURBATION THEORY FOR ALGEBRAIC RICCATI EQUATIONS*

JI-GUANG SUN†

Abstract. New perturbation results for the two different algebraic Riccati equations (continuous time and discrete time) are derived in a uniform manner. The new results are illustrated by numerical examples.

Key words. algebraic Riccati equations, Hermitian positive semidefinite solution, perturbation bound, condition number

AMS subject classifications. 15A24, 65H05, 93B35

PII. S0895479895291303

1. Introduction.

We consider the continuous-time algebraic Riccati equation (CARE)

$$Q + A^H X + XA - XBR^{-1}B^H X = 0 \tag{1.1}$$

and the discrete-time algebraic Riccati equation (DARE)

$$X - A^H XA + A^H XB(R + B^H XB)^{-1}B^H XA - C^H C = 0. \tag{1.2}$$

Appropriate assumptions on the coefficient matrices will be made in sections 1.1 and 1.2 to guarantee the existence and uniqueness of the Hermitian positive semidefinite (p.s.d.) solution. Equations (1.1) and (1.2) arise naturally in linear control and system theory, and there are many contributions in the literature on the theory, applications, and numerical solution of the equations (see, e.g., [1], [17], [22], [23]). Although for applications the real case, i.e., when all the coefficient matrices are real and real solution matrices X are to be found, is especially important, we consider here the general, i.e., complex, case as well as the real case.

The central question of perturbation theory for an algebraic Riccati equation is as follows: How does the Hermitian p.s.d. solution X change when the coefficient matrices are subject to perturbations? The interest in this topic is motivated by the fact that these equations are usually subject to perturbations in the coefficient matrices reflecting various errors in the formulation of the problems and in their solutions by a computer. (See, e.g., [2], [11], [18], [19], [21], [22], [29] for numerical methods for solving the equations.)

Perturbation theory for the algebraic Riccati equations (1.1) and (1.2) are studied by a number of authors [3], [5], [7], [10], [12], [13], [14], [15], [16], [26], [27], [28], [32]. This paper, as a continuation of the previous results of other authors [3], [5], [14], [15], [16], [32], derives new perturbation bounds for the Hermitian p.s.d. solution to the CARE (1.1) and for the Hermitian p.s.d. solution to the DARE (1.2) in a uniform manner. The new results are illustrated by numerical examples.

*Received by the editors September 1, 1995; accepted for publication (in revised form) by P. van Dooren December 6, 1996. This work was supported by the Swedish Natural Science Research Council under contract M-AA/MA 06952-303 and the Department of Computing Science, UmeåUniversity, Umeå, Sweden.

http://www.siam.org/journals/simax/19-1/29130.html

†Department of Computing Science, UmeåUniversity, S-901 87 Umeå, Sweden (jisun@cs.umu.se).

Throughout this paper we use $\mathcal{C}^{m\times n}$ (or $\mathcal{R}^{m\times n}$) to denote the set of complex (or real) $m\times n$ matrices, and $\mathcal{H}^{n\times n}$ to denote the set of $n\times n$ Hermitian matrices. $\bar{A}$ denotes the conjugate of a matrix A, A^T denotes the transpose of A, and $A^H=\bar{A}^T$. I stands for the identity matrix, I_n is the identity matrix of order n, and 0 is the null matrix. The positive definiteness (or semidefiniteness) of a Hermitian matrix A will be denoted by $A>0$ (or $A\ge 0$). The eigenvalues of $A\in\mathcal{C}^{n\times n}$ are denoted by $\lambda_1(A),\lambda_2(A),\ldots,\lambda_n(A)$, and $\lambda(A)=\{\lambda_i(A)\}_{i=1}^n$. The spectral radius $\rho(A)$ is defined by $\rho(A)=\max_i|\lambda_i(A)|$. The symbol $\|\ \|$ stands for any unitarily invariant norm, $\|\ \|_F$ is the Frobenius norm, and $\|\ \|_2$ is the spectral norm. For $A=(a_1,\ldots,a_n)=(\alpha_{ij})\in\mathcal{C}^{m\times n}$ and a matrix B, $A\otimes B=(\alpha_{ij}B)$ is a Kronecker product, and $\mathrm{vec}(A)$ is a vector defined by $\mathrm{vec}(A)=(a_1^T,\ldots,a_n^T)^T$. (See [9, Chaps. 1 and 2] for properties of the Kronecker product and vec operation.)

The stable matrix is an important notion to the study of the CARE and DARE. An $n\times n$ matrix A is said to be c-stable if all the eigenvalues of A lie in the open left-half complex plane, and A is said to be d-stable if $\rho(A)<1$.

1.1. Problem statement of the CARE. The coefficient matrices of the CARE (1.1) are $A\in\mathcal{C}^{n\times n}$, $B\in\mathcal{C}^{n\times m}$, $Q\in\mathcal{H}^{n\times n}$, and $R\in\mathcal{H}^{m\times m}$, in which $Q\ge 0$ and $R>0$. Let $G=BR^{-1}B^H$. Then the CARE (1.1) can be written in the equivalent form

$$Q+A^HX+XA-XGX=0, \tag{1.3}$$

where $Q,G\ge 0$.

Throughout this paper we assume that (A,G) is a c-stabilizable pair, i.e., there is a matrix $K\in\mathcal{C}^{n\times n}$ such that the matrix $A-GK$ is c-stable, and that (A,Q) is a c-detectable pair, i.e., if (A^T,Q^T) is c-stabilizable. It is known [3], [18] that in such a case there exists a unique Hermitian p.s.d. solution X to the CARE (1.3), and the matrix $A-GX$ is c-stable.

Many perturbation results for the CARE (1.3) can be found in the literature. Byers [3], and Kenney and Hewer [12] obtain the first-order perturbation bounds for the solution to the CARE (1.3). Chen [5] and Konstantinov, Petkov, and Christov [14] derive global perturbation bounds for the solution. Xu [32] improves Chen's results, and Konstantinov, Petkov, Gu, and Postlethwaite [16] sharpen the results of [14]. Kenney, Laub, and Wette [13] derive residual error bounds associated with Newton refinement of approximate solutions. Ghavimi and Laub [7] present a new backward error criterion, together with a sensitivity measure, for assessing solution accuracy. Besides, a new residual bound for an approximate solution to (1.3) is presented in [26].

Let X be the unique Hermitian p.s.d. solution to the CARE (1.3). Under some hypotheses, Konstantinov, Petkov, and Christov [14] and Konstantinov, Petkov, Gu, and Postlethwaite [16] derive upper bounds for $\|\tilde{X}-X\|_F$, where $\tilde{X}$ is a solution to the perturbed CARE

$$\tilde{Q}+\tilde{A}^H\tilde{X}+\tilde{X}\tilde{A}-\tilde{X}\tilde{G}\tilde{X}=0. \tag{1.4}$$

However, it is not considered whether the solution $\tilde{X}$ is Hermitian p.s.d., and it is not even considered whether the perturbed equation (1.3) has a Hermitian p.s.d. solution. Recently, Xu [32] described a technique for discussing perturbations of the Hermitian p.s.d. solution X of the CARE (1.3), and presented an upper bound for $\|\tilde{X}-X\|_C/\|X\|_C$, where $\tilde{X}$ is a Hermitian p.s.d. solution to the perturbed CARE

(1.4), and $\|\ \|_C$ denotes any consistent norm on $\mathcal{C}^{n\times n}$ with $\|I\|_C = 1$, but the new upper bound given by [32] is as conservative as that of [14].

In section 3 we shall present some reasonable restrictions on the perturbations in the coefficient matrices and derive a sharp upper bound for $\|\tilde{X} - X\|$, where $\tilde{X}$ is the unique Hermitian p.s.d. solution to the perturbed CARE (1.4). The first-order perturbation bound and condition numbers of the unique Hermitian p.s.d. solution X to the CARE (1.3) are then deduced from the new perturbation result.

1.2. Problem statement of the DARE. The coefficient matrices of the DARE (1.2) are $A \in \mathcal{C}^{n\times n}$, $B \in \mathcal{C}^{n\times m}$, $C \in \mathcal{C}^{r\times n}$, and $R \in \mathcal{C}^{m\times m}$. As usual, we assume $R^H = R > 0$. (Note that the perturbation analysis of a more general DARE, where the assumption $R > 0$ is dropped [11], will be studied separately.) Let $G = BR^{-1}B^H$, $Q = C^HC$. Then the DARE (1.2) can be written in the equivalent form

$$X - A^HX(I + GX)^{-1}A - Q = 0, \tag{1.5}$$

where $Q, G \geq 0$.

Throughout this paper we assume that (A, B) is a d-stabilizable pair, i.e., if $w^HB = 0$ and $w^HA = \lambda w^H$ for some constant λ implies $|\lambda| < 1$ or $w = 0$, and that (A, C) is a d-detectable pair, i.e., if (A^T, C^T) is d-stabilizable [21]. It is known [1], [10], [15] that in such a case there exists a unique Hermitian p.s.d. solution X to the DARE (1.5), and the matrix $(I + GX)^{-1}A$ is d-stable.

Perturbation theory for the DARE (1.5) is studied by a certain number of authors. Gudmundsson, Kenney, and Laub [10] derive a condition number of the DARE (1.5) and a bound on the relative error of a computed solution. Konstantinov, Petkov, and Christov [15] obtain perturbation bounds and determine the conditioning of the equation. Computable residual bounds of an approximate solution to the DARE (1.5) are derived by [27], and the normwise backward error of an approximate solution is evaluated by [28].

Let X be the unique Hermitian p.s.d. solution to the DARE (1.5). Under some hypotheses, Konstantinov, Petkov, and Christov [15] derive upper bounds for $\|\tilde{X} - X\|_F$ and $\|\tilde{X} - X\|_2$, where $\tilde{X}$ is a Hermitian solution to the perturbed DARE

$$\tilde{X} - \tilde{A}^H\tilde{X}(I + \tilde{G}\tilde{X})^{-1}\tilde{A} - \tilde{Q} = 0. \tag{1.6}$$

However, it is not considered whether the solution $\tilde{X}$ is p.s.d., and it is not even considered whether the perturbed DARE (1.6) has a Hermitian p.s.d. solution. Moreover, the upper bounds given by [15] can be improved.

One of the difficult points for deriving a sharp upper bound for $\|\tilde{X} - X\|$ is how to find an equation of $\Delta X \equiv \tilde{X} - X$ which is easy to handle. Another difficult point is how to find some reasonable restrictions on the perturbations in the coefficient matrices of the DARE (1.5) such that the perturbed DARE (1.6) has a unique Hermitian p.s.d. solution $\tilde{X}$, and that it is easy to estimate $\|\tilde{X} - X\|$.

In section 4, we shall present some reasonable restrictions on the perturbations in the coefficient matrices and derive a new upper bound for $\|\tilde{X} - X\|$, where $\tilde{X}$ is the unique Hermitian p.s.d. solution to the perturbed DARE (1.6). The first-order perturbation bound and condition numbers of the unique Hermitian p.s.d. solution X to the DARE (1.5) are then deduced from the new perturbation result.

The rest of this paper is organized as follows. We begin in section 2 with some lemmas on perturbation properties of the stable matrices and on the uniqueness of

the stabilizing solution. In sections 3 and 4 we derive new perturbation results for the CARE and DARE, respectively. The new results will be illustrated by simple numerical examples in section 5. Finally, in the Appendix we give a proof of several useful formulae.

2. Lemmas.

2.1. Perturbation properties of stable matrices. Let $\Phi \in \mathcal{C}^{n\times n}$. Define the linear operator $\mathbf{L}_c$: $\mathcal{H}^{n\times n} \to \mathcal{H}^{n\times n}$ by

$$\mathbf{L}_c W = \Phi^H W + W\Phi, \qquad W \in \mathcal{H}^{n\times n}. \tag{2.1}$$

It is known (see, e.g., [25, pp. 222–223]) that if Φ is c-stable, then $\mathbf{L}_c$ is invertible. The following lemma will be used in section 3.

LEMMA 2.1 (see [26, Corollary 2.5]). *Let $\mathbf{L}_c$ be the linear operator defined by* (2.1) *with a c-stable matrix $\Phi \in \mathcal{C}^{n\times n}$. If $E \in \mathcal{C}^{n\times n}$ satisfies*

$$2\|\mathbf{L}_c^{-1}\|\|E\| < 1,$$

then $\Phi + E$ is c-stable.

Let $\Phi \in \mathcal{C}^{n\times n}$. Define the linear operator $\mathbf{L}_d$: $\mathcal{H}^{n\times n} \to \mathcal{H}^{n\times n}$ by

$$\mathbf{L}_d W = W - \Phi^H W\Phi, \qquad W \in \mathcal{H}^{n\times n}. \tag{2.2}$$

It is known [6] that if Φ is d-stable, then $\mathbf{L}_d$ is invertible. The following lemma will be used in section 4.

LEMMA 2.2. *Let $\mathbf{L}_d$ be the linear operator defined by* (2.2) *with a d-stable matrix $\Phi \in \mathcal{C}^{n\times n}$, and let*

$$l_d = \|\mathbf{L}_d^{-1}\|^{-1}, \qquad \phi = \|\Phi\|_2. \tag{2.3}$$

If $E \in \mathcal{C}^{n\times n}$ satisfies

$$\|E\| < \frac{l_d}{\phi + \sqrt{\phi^2 + l_d}}, \tag{2.4}$$

then $\Phi + E$ is d-stable.

Lemma 2.2 is a corollary of Lemma 2.4 of this section. We first prove the following lemma.

LEMMA 2.3. *Let $\Phi \in \mathcal{C}^{n\times n}$. If there is an eigenvalue $\phi_k \in \lambda(\Phi)$ with $|\phi_k| = 1$, then the operator $\mathbf{L}_d$ defined by* (2.2) *is singular.*

Proof. We only need to prove that under the hypothesis there is a Hermitian matrix $W \neq 0$ such that $H(W) \equiv W - \Phi^H W\Phi = 0$.

Let $\Phi^H = UTU^H$ be the Schur decomposition of Φ^H, where U is unitary, and T is upper triangular with the diagonal elements $\phi_1, \ldots, \phi_n$, the eigenvalues of Φ (see [8, Chap. 7] for the Schur decomposition). Without loss of generality we may assume that $\Phi^H = T$ and $k = 2$, i.e., Φ^H has the form

$$\Phi^H = \begin{pmatrix} T_1 & * \\ 0 & * \end{pmatrix} \quad \text{with} \quad T_1 = \begin{pmatrix} \phi_1 & t \\ 0 & \phi_2 \end{pmatrix}, \ \ |\phi_2| = 1.$$

Observe that if we take

$$W = \begin{pmatrix} W_1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{with} \quad W_1 = \begin{pmatrix} \omega_1 & w \\ \bar{w} & \omega_2 \end{pmatrix},$$

then

$$H(W) = \begin{pmatrix} H_1(W_1) & 0 \\ 0 & 0 \end{pmatrix} \quad \text{with} \quad H_1(W_1) = W_1 - T_1 W_1 T_1^H = \begin{pmatrix} \eta_1 & h \\ \bar{h} & \eta_2 \end{pmatrix},$$

where

$$\begin{aligned} &\eta_1 = (1-|\phi_1|^2)\omega_1 - |t|^2\omega_2 - \phi_1\bar{t}w - \bar{\phi}_1 t\bar{w}, \\ &\eta_2 = (1-|\phi_2|^2)\omega_2 = 0, \qquad h = -\bar{\phi}_1 t\omega_2 + (1-\phi_1\bar{\phi}_2)w. \end{aligned} \tag{2.5}$$

There are two possibilities: (a) $|\phi_1| \neq 1$ or (b) $|\phi_1| = 1$. In the former case (a), we can take $\omega_2 = 1$, and it is easy to verify that by (2.5) the Hermitian matrix

$$W_1 = \begin{pmatrix} \omega_1^* & w^* \\ \bar{w}^* & 1 \end{pmatrix} \quad \text{with} \quad w^* = \frac{\bar{\phi}_1 t}{1-\phi_1\bar{\phi}_2}, \quad \omega_1^* = \frac{|t|^2 + \phi_1\bar{t}w^* + \bar{\phi}_1 t\bar{w}^*}{1-|\phi_1|^2}$$

satisfies $W_1 \neq 0$ and $H_1(W_1) = 0$, thereby there is a Hermitian matrix $W = \begin{pmatrix} W_1 & 0 \\ 0 & 0 \end{pmatrix} \neq 0$ such that $H(W) = 0$. In the latter case (b), we can take $\omega_1 = 1$, and it is evident that by (2.5) the Hermitian matrix $W_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ satisfies $W_1 \neq 0$ and $H_1(W_1) = 0$, thereby there is a Hermitian matrix $W = \begin{pmatrix} W_1 & 0 \\ 0 & 0 \end{pmatrix} \neq 0$ such that $H(W) = 0$. The proof is completed. □

Referring to [30], for the d-stable matrix Φ we define the quantity s by

$$s = \min\left\{ \|E\| : \max_{1\le j\le n} |\lambda_j(\Phi+E)| \ge 1,\ E \in \mathcal{C}^{n\times n} \right\}.$$

The quantity s measures the size of the smallest $\|E\|$ such that $\Phi+E$ has an eigenvalue λ_{j^*} with $|\lambda_{j^*}| \ge 1$. By the continuity of the eigenvalues we have

$$s = \min\left\{ \|E\| : \max_{1\le j\le n} |\lambda_j(\Phi+E)| = 1,\ E \in \mathcal{C}^{n\times n} \right\}. \tag{2.6}$$

This means that the quantity s measures the smallest $\|E\|$ such that $\Phi + E$ has an eigenvalue on the unit circle.

The following result establishes a connection between the quantities s, l_d, and $\rho(\Phi)$.

LEMMA 2.4. *Let $\Phi \in \mathcal{C}^{n\times n}$ be d-stable, and let l_d and s be defined by* (2.3) *and* (2.6)*, respectively. Moreover, let $\rho = \rho(\Phi)$ and $\phi = \|\Phi\|_2$. Then*

$$\frac{l_d}{\phi + \sqrt{\phi^2 + l_d}} \le s \le 1 - \rho. \tag{2.7}$$

Proof. Let $E_* \in \mathcal{C}^{n\times n}$ be such that

$$s(\Phi) = \|E_*\| \quad \text{with} \quad \max_{1\le j\le n} |\lambda_j(\Phi+E_*)| = 1.$$

Then by Lemma 2.3, the transformation

$$W \to W - (\Phi+E_*)^H W(\Phi+E_*) \quad \text{with} \quad W \in \mathcal{H}^{n\times n}$$

is singular, i.e., there is a Hermitian matrix $W_* \neq 0$ such that

$$W_* - (\Phi + E_*)^H W_* (\Phi + E_*) = 0,$$

or, equivalently,

$$\mathbf{L}_d W_* = \Phi^H W_* E_* + E_*^H W_* \Phi + E_*^H W_* E_*, \tag{2.8}$$

where $\mathbf{L}_d$ is defined by (2.2). From (2.8)

$$s^2 + 2\phi s - l_d \geq 0,$$

which implies the first inequality of (2.7).

Let $\Phi = UTU^H$ be the Schur decomposition of Φ, where U is unitary, and T is upper triangular with the diagonal elements $\lambda_1(\Phi), \ldots, \lambda_n(\Phi)$. Assume that $\rho(\Phi) = |\lambda_{i^*}(\Phi)|$. By (2.6) we have

$$\begin{aligned} s &= \min\left\{ \|E\| : \max_{1\leq j\leq n} \left|\lambda_j(T + U^H E U)\right| = 1 \right\} \\ &\leq \min\left\{ \|D\| : \max_{1\leq j\leq n} |\lambda_j(T + D)| = 1 \right\}, \end{aligned} \tag{2.9}$$

where $D = \mathrm{diag}(\delta_i)$ with complex scalars $\delta_i, i = 1, \ldots, n$. Take

$$\delta_i = \begin{cases} -\lambda_i(\Phi) + \mathrm{e}^{i\arg(\lambda_i(\Phi))} & \text{if } i = i^*, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\max_{1\leq j\leq n} |\lambda_j(T + D)| = 1 \quad \text{with} \quad \|D\| = 1 - \rho.$$

Substituting it into (2.9) gives the second inequality of (2.7). ∎

From Lemma 2.4 we get Lemma 2.2 immediately.

2.2. The uniqueness of the stabilizing solution. We call $X \in \mathcal{H}^{n\times n}$ a c-stabilizing solution to the CARE (1.3) if X satisfies (1.3), and $A - GX$ is c-stable. Similarly, we call $X \in \mathcal{H}^{n\times n}$ a d-stabilizing solution to the DARE (1.5) if X satisfies (1.5), and $(I + GX)^{-1}A$ is d-stable.

LEMMA 2.5. *If the CARE* (1.3) *has a c-stabilizing solution, then it is unique.*

Proof. Let X_1 and X_2 be two c-stabilizing solutions to the CARE (1.3). Then from

$$Q + A^H X_i + X_i A - X_i G X_i = 0, \quad i = 1, 2,$$

we get

$$(A - GX_2)^H (X_1 - X_2) + (X_1 - X_2)(A - GX_1) = 0. \tag{2.10}$$

Since both $A - GX_1$ and $A - GX_2$ are c-stable, (2.10) has the unique solution $X_1 - X_2 = 0$, i.e., $X_1 = X_2$. ∎

LEMMA 2.6 (see [11, Proposition 1]). *If the DARE* (1.5) *has a d-stabilizing solution, then it is unique.*

3. Perturbation results for the CARE.

3.1. Perturbation equation. Let X be the unique Hermitian p.s.d. solution to the CARE (1.3), and let $\tilde{X}$ be a Hermitian solution to the perturbed CARE (1.4). Define

$$\Delta X = \tilde{X} - X, \quad \Delta Q = \tilde{Q} - Q, \quad \Delta A = \tilde{A} - A, \quad \Delta G = \tilde{G} - G.$$

Then from (1.3) and (1.4) we see that the matrix ΔX satisfies the equation

$$(A - GX)^H \Delta X + \Delta X(A - GX) = -E + h_1(\Delta X) + h_2(\Delta X), \tag{3.1}$$

where

$$E = \Delta Q + \Delta A^H X + X\Delta A - X\Delta G X, \tag{3.2}$$

and

$$\begin{aligned} h_1(\Delta X) &= -\left[(\Delta A - \Delta G X)^H \Delta X + \Delta X(\Delta A - \Delta G X)\right], \\ h_2(\Delta X) &= \Delta X(G + \Delta G)\Delta X. \end{aligned} \tag{3.3}$$

Let $\Phi = A - GX$, and define the linear operator $\mathbf{L}$: $\mathcal{H}^{n\times n} \to \mathcal{H}^{n\times n}$ by

$$\mathbf{L}W = \Phi^H W + W\Phi, \quad W \in \mathcal{H}^{n\times n}. \tag{3.4}$$

Then (3.1) can be written as

$$\mathbf{L}\Delta X = -E + h_1(\Delta X) + h_2(\Delta X). \tag{3.5}$$

Since Φ is c-stable, the operator $\mathbf{L}$ is invertible. Define the function $\mu(\Delta X)$ by

$$\mu(\Delta X) = -\mathbf{L}^{-1}E + \mathbf{L}^{-1}[h_1(\Delta X) + h_2(\Delta X)]. \tag{3.6}$$

Obviously, $\mu(\Delta X)$ can be regarded as a continuous mapping $\mathcal{M} : \mathcal{H}^{n\times n} \to \mathcal{H}^{n\times n}$, and the set of the solutions to (3.5) is just the set of the fixed points of the mapping $\mathcal{M}$.

Define the linear operators $\mathbf{P}$: $\mathcal{C}^{n\times n} \to \mathcal{H}^{n\times n}$, and $\mathbf{Q}$: $\mathcal{H}^{n\times n} \to \mathcal{H}^{n\times n}$ by [3]

$$\mathbf{P}N = \mathbf{L}^{-1}(XN + N^H X), \quad N \in \mathcal{C}^{n\times n}, \tag{3.7}$$

and

$$\mathbf{Q}H = \mathbf{L}^{-1}(XHX), \quad H \in \mathcal{H}^{n\times n}. \tag{3.8}$$

Then by (3.2) we have

$$\mathbf{L}^{-1}E = \mathbf{L}^{-1}\Delta Q + \mathbf{P}\Delta A - \mathbf{Q}\Delta G. \tag{3.9}$$

In the following subsections we derive an upper bound for some fixed points ΔX of the continuous mapping $\mathcal{M}$ expressed by (3.6) under some assumptions on $\Delta Q, \Delta A$, and ΔG, where $\Delta X = \tilde{X} - X$, in which X is the unique Hermitian p.s.d. solution to the CARE (1.3), and $\tilde{X}$ is the unique Hermitian p.s.d. solution to the perturbed CARE (1.4). For simplicity, we assume that $\tilde{Q}, \tilde{G} \geq 0$.

3.2. Estimates of some fixed points of $\mathcal{M}$. Let $\mathbf{L}$, $\mathbf{P}$, $\mathbf{Q}$ be the linear operators defined by (3.4), (3.7), and (3.8), respectively. Define

$$\begin{aligned} &l = \|\mathbf{L}^{-1}\|^{-1}, \quad p = \|\mathbf{P}\|, \quad q = \|\mathbf{Q}\|, \\ &\delta = \|\Delta A\| + \|\Delta G\|_2\|X\|, \quad g = \|G\|_2, \quad \hat{g} = g + \|\Delta G\|_2. \end{aligned} \tag{3.10}$$

Observe that from (3.3) and (3.10)

$$\|h_1(\Delta X)\| \le 2\delta\|\Delta X\|, \quad \|h_2(\Delta X)\| \le \hat{g}\|\Delta X\|^2. \tag{3.11}$$

Hence, by (3.5), (3.9), (3.10), and (3.11), ΔX satisfies

$$\|\Delta X\| \le \epsilon + \frac{2\delta}{l}\|\Delta X\| + \frac{\hat{g}}{l}\|\Delta X\|^2, \quad \text{i.e.,} \quad \hat{g}\|\Delta X\|^2 - (l-2\delta)\|\Delta X\| + l\epsilon \ge 0, \tag{3.12}$$

where ϵ is defined by

$$\epsilon = \frac{1}{l}\|\Delta Q\| + p\|\Delta A\| + q\|\Delta G\|.$$

Consider the equation

$$\hat{g}\xi^2 - (l-2\delta)\xi + l\epsilon = 0. \tag{3.13}$$

It can be verified that if δ and ϵ satisfy

$$\delta < \frac{l}{2} \quad \text{and} \quad \epsilon \le \frac{(l-2\delta)^2}{4l\hat{g}}, \tag{3.14}$$

then the positive scalar ξ_* expressed by

$$\xi_* = \frac{2l\epsilon}{l - 2\delta + \sqrt{(l-2\delta)^2 - 4l\hat{g}\epsilon}} \tag{3.15}$$

is a solution to (3.13), and the function $\mu(\Delta X)$ defined by (3.6) satisfies

$$\|\mu(\Delta X)\| \le \epsilon + \frac{2\delta}{l}\|\Delta X\| + \frac{\hat{g}}{l}\|\Delta X\|^2 \le \xi_* \quad \text{if} \quad \|\Delta X\| \le \xi_*. \tag{3.16}$$

It is known that the space $\mathcal{H}^{n\times n}$ with any unitarily invariant norm $\|\ \|$ is a Banach space. We now consider the set $\mathcal{S}_{\xi_*} \subset \mathcal{H}^{n\times n}$ defined by

$$\mathcal{S}_{\xi_*} = \{\Delta X \in \mathcal{H}^{n\times n} \ : \ \|\Delta X\| \le \xi_*\}.$$

Since $\mathcal{S}_{\xi_*}$ is a bounded closed convex set of $\mathcal{H}^{n\times n}$, and the relation (3.16) shows that the continuous mapping $\mathcal{M}$ expressed by (3.6) maps $\mathcal{S}_{\xi_*}$ into $\mathcal{S}_{\xi_*}$, by the Schauder fixed-point theorem (see, e.g., [20, sect. 6.3]), the mapping $\mathcal{M}$ has a fixed point $\Delta X_* \in \mathcal{S}_{\xi_*}$, i.e.,

$$\|\Delta X_*\| \le \xi_*, \tag{3.17}$$

where ξ_* is expressed by (3.15).

3.3. The matrix $X + \Delta X_*$. Let $\Delta X_* \in \mathcal{S}_{\xi_*}$ be a fixed point of the mapping $\mathcal{M}$ expressed by (3.6). Define $Y = X + \Delta X_*$. Then from section 3.1 we see that the Hermitian matrix Y satisfies

$$\tilde{Q} + \tilde{A}^H Y + Y\tilde{A} - Y\tilde{G}Y = 0, \tag{3.18}$$

i.e., Y is a Hermitian solution to the perturbed CARE (1.4).

It is easy to verify that (3.18) can be written as

$$(\tilde{A} - \tilde{G}Y)^H Y + Y(\tilde{A} - \tilde{G}Y) = -(\tilde{Q} + Y\tilde{G}Y). \tag{3.19}$$

Observe the following facts: the matrix $\tilde{Q}+Y\tilde{G}Y$ is Hermitian p.s.d.; from $\tilde{A} = A+\Delta A$ and $Y = X + \Delta X_*$ we have

$$\tilde{A} - \tilde{G}Y = \Phi + \Delta A - \Delta G X - (G + \Delta G)\Delta X_*,$$

where $\Phi = A - GX$ is stable, and from (3.10) and (3.17)

$$\|\Delta A - \Delta G X - (G + \Delta G)\Delta X_*\| \le \delta + \hat{g}\xi_*.$$

Consequently, by Lemma 2.1, if

$$2(\delta + \hat{g}\xi_*)/l < 1, \quad \text{i.e.,} \quad \delta + \hat{g}\xi_* < l/2, \tag{3.20}$$

then the matrix $\tilde{A} - \tilde{G}Y$ is also stable. Hence, by Lemma 2.5, under the condition (3.20) the matrix Y, as a c-stabilizing solution to the CARE (3.18), is unique. Moreover, the Hermitian matrix Y, as a solution to (3.19), is p.s.d. [31, Lemma 12.1].

Thus, we have proved that under conditions (3.14) and (3.20), there is a unique Hermitian p.s.d. solution $\tilde{X} = Y$ to the CARE (1.4), and $\|\tilde{X} - X\| \le \xi_*$, where X is the unique Hermitian p.s.d. solution to the CARE (1.3).

3.4. The conditions (3.14) and (3.20). Note that the condition (3.20) can be deduced from the conditions (3.14). In fact, the conditions (3.14) imply that the positive scalar ξ_* expressed by (3.15) satisfies

$$\xi_* \le (l - 2\delta)/(2\hat{g}),$$

which is just the condition (3.20). Moreover, the conditions (3.14) can be expressed by an equivalent condition

$$\delta + \sqrt{l\hat{g}\epsilon} < \frac{l}{2}.$$

3.5. Perturbation bounds. Overall, we have the following theorem.

THEOREM 3.1. *Let X be the unique Hermitian p.s.d. solution to the CARE* (1.3)*. Define the linear operators* $\mathbf{L}$, $\mathbf{P}$, *and* $\mathbf{Q}$ *by* (3.4), (3.7), *and* (3.8), *and define* $l, p, q, \delta, \hat{g}$ *by* (3.10), *respectively. Moreover, let* $\tilde{Q} = Q + \Delta Q$, $\tilde{A} = A + \Delta A$, *and* $\tilde{G} = G + \Delta G$ *be the coefficient matrices of the perturbed CARE* (1.4), *and let*

$$\epsilon = \frac{1}{l}\|\Delta Q\| + p\|\Delta A\| + q\|\Delta G\|, \tag{3.21}$$

and

$$\xi_* = \frac{2l\epsilon}{l - 2\delta + \sqrt{(l - 2\delta)^2 - 4l\hat{g}\epsilon}}. \tag{3.22}$$

If $\tilde{G}, \tilde{Q} \geq 0$*, and if*

$$\delta + \sqrt{l\hat{g}\epsilon} < \frac{l}{2}, \tag{3.23}$$

then the CARE (1.4) *has a unique Hermitian p.s.d. solution* $\tilde{X}$*, and*

$$\|\tilde{X} - X\| \leq \xi_*. \tag{3.24}$$

From Theorem 3.1 we get the first-order perturbation bound for the solution X:

$$\|\tilde{X} - X\| \leq \epsilon + O(\|(\Delta Q, \Delta A, \Delta G)\|^2), \qquad \|(\Delta Q, \Delta A, \Delta G)\| \to 0, \tag{3.25}$$

where ϵ is defined by (3.21). Consequently, for sufficiently small $\|(\Delta Q, \Delta A, \Delta G)\|$, we have

$$\frac{\|\tilde{X} - X\|}{\|X\|} \lesssim \frac{\|Q\|}{l\|X\|}\frac{\|\Delta Q\|}{\|Q\|} + \frac{p\|A\|}{\|X\|}\frac{\|\Delta A\|}{\|A\|} + \frac{q\|G\|}{\|X\|}\frac{\|\Delta G\|}{\|G\|}. \tag{3.26}$$

Remark 3.2. Konstantinov et al. [16, sect. 3.2.1] present the same upper bound ξ_* for $\|\tilde{X} - X\|_F$, where ξ_* is expressed by (3.22), in which ϵ is defined by (3.21) in the Frobenius norm $\|\ \|_F$. However, [16, sect. 3.2.1] does not distinguish whether or not the solution $\tilde{X}$ to the perturbed CARE (1.4) is Hermitian, and it does not even know whether or not (1.4) has a Hermitian p.s.d. solution. Recently, Xu [32] presented an upper bound for $\|\tilde{X} - X\|_C/\|X\|_C$, where $\tilde{X}$ is a Hermitian p.s.d. solution to the CARE (1.4), and $\|\ \|_C$ denotes any consistent norm on $\mathcal{C}^{n\times n}$ with $\|I\|_C = 1$. But the bound obtained by [32] is conservative. For comparing our result (Theorem 3.1) with that of [32], we now take the spectral norm $\|\ \|_2$. By [32, Thm. 2.1], for sufficiently small $\|(\Delta Q, \Delta A, \Delta G)\|_2$ we have

$$\frac{\|\tilde{X} - X\|_2}{\|X\|_2} \lesssim \frac{2}{l_2}(2\|A\|_2 + \|X\|_2\|G\|_2)\left(\frac{\|\Delta A\|_2}{\|A\|_2} + \frac{\|\Delta G\|_2}{\|G\|_2} + \frac{\|\Delta Q\|_2}{\|Q\|_2}\right) \equiv \chi, \tag{3.27}$$

where $l_2 = \|\mathbf{L}^{-1}\|^{-1}$, in which the operator norm $\|\ \|$ is induced by the spectral norm $\|\ \|_2$. By Theorem 3.1 we have the estimate (3.26). Observe that from (1.3), (3.10), (3.7), and (3.8)

$$\|Q\|_2 \leq \|X\|_2(2\|A\|_2 + \|X\|_2\|G\|_2), \qquad p \leq 2\|X\|_2/l_2, \qquad q \leq \|X\|_2^2/l_2.$$

Hence, the estimate (3.26) implies

$$\begin{aligned}\frac{\|\tilde{X} - X\|_2}{\|X\|_2} \lesssim\ & \tfrac{1}{l_2}\Big[2\|A\|_2\tfrac{\|\Delta A\|_2}{\|A\|_2} + \|G\|_2\|X\|_2\tfrac{\|\Delta G\|_2}{\|G\|_2} \\ & + (2\|A\|_2 + \|X\|_2\|G\|_2)\frac{\|\Delta Q\|_2}{\|Q\|_2}\Big] \leq \chi/2,\end{aligned}$$

where χ is defined by (3.27). Consequently, the result of Theorem 3.1 is better than that of [32].

Note that the quantity l_2 is difficult to compute. Suppose that the operator norm $\|\ \|$ in the definition $l = \|\mathbf{L}^{-1}\|^{-1}$ is induced by the Frobenius norm $\|\ \|_F$ (see section

3.7 for the computation of the scalar l). From $l/\sqrt{n} \le l_2 \le \sqrt{n}l$ [24, Thm. 4.10] we see that if we define $\hat{\chi}$ by

$$\hat{\chi} = \frac{2}{l}(2\|A\|_2 + \|X\|_2\|G\|_2)\left(\frac{\|\Delta A\|_2}{\|A\|_2} + \frac{\|\Delta G\|_2}{\|G\|_2} + \frac{\|\Delta Q\|_2}{\|Q\|_2}\right),$$

then

$$\chi^{(l)} \equiv \frac{1}{\sqrt{n}}\hat{\chi} \le \chi \le \sqrt{n}\hat{\chi} \equiv \chi^{(u)}. \tag{3.28}$$

3.6. Condition numbers. The relation (3.26) shows that the scalars $c_Q(X)$, $c_A(X)$, $c_G(X)$, and $c_Q^{(r)}(X)$, $c_A^{(r)}(X)$, $c_G^{(r)}(X)$, defined by

$$c_Q(X) = \frac{1}{l}, \quad c_A(X) = p, \quad c_G(X) = q$$

and

$$c_Q^{(r)}(X) = \frac{\|Q\|}{l\|X\|}, \quad c_A^{(r)}(X) = \frac{p\|A\|}{\|X\|}, \quad c_G^{(r)}(X) = \frac{q\|G\|}{\|X\|},$$

are the absolute and relative condition numbers of X with respect to Q, A, G, respectively. Moreover, the scalar $c^{(r)}(X)$ defined by

$$c^{(r)}(X) = \frac{1}{\|X\|}\sqrt{(\|Q\|/l)^2 + (p\|A\|)^2 + (q\|G\|)^2} \tag{3.29}$$

can be regarded as the relative condition number of X.

By using a local linear estimate, Byers [3] presents a condition number $\kappa_B(X)$ of X:

$$\kappa_B(X) = \frac{1}{\|X\|_F}\left(\|Q\|_F/l + p\|A\|_F + q\|G\|_F\right), \tag{3.30}$$

in which the operator norm $\|\ \|$ for defining l, p, q is induced by the Frobenius norm $\|\ \|_F$. We now take the Frobenius norm in (3.29). Comparing (3.29) with (3.30) gives

$$\frac{1}{\sqrt{3}}\kappa_B(X) \le c^{(r)}(X) \le \kappa_B(X).$$

3.7. Expressions of $\|\mathbf{L}^{-1}\|^{-1}$, $\|\mathbf{P}\|$, and $\|\mathbf{Q}\|$. Let $\mathbf{L}$, $\mathbf{P}$, and $\mathbf{Q}$ be the linear operators defined by (3.4), (3.7), and (3.8), respectively, and let $l = \|\mathbf{L}^{-1}\|^{-1}$, $p = \|\mathbf{P}\|$, and $q = \|\mathbf{Q}\|$ (see (3.10)). The problem of finding explicit expressions of l, p, q is a difficult one. However, if the operator norm $\|\ \|$ for defining l and q is induced by the Frobenius norm on $\mathcal{H}^{n\times n}$, and the operator norm $\|\ \|$ for defining p is induced by the Frobenius norm $\|\ \|_F$ on $\mathcal{C}^{n\times n}$, then by using the technique described by Byers and Nash [4] we can find explicit expressions of the corresponding l, p, and q: define the matrix T by

$$T = I_n \otimes \Phi^H + \Phi^T \otimes I_n, \tag{3.31}$$

where $\Phi = A - GX$ is c-stable. Then

$$l = \|T^{-1}\|_2^{-1}, \tag{3.32}$$

$$p = \|T^{-1}[(I_n \otimes X)(I_{n^2},\, iI_{n^2}) + (X^T \otimes I_n)\Pi(I_{n^2},\, -iI_{n^2})]\|_2, \tag{3.33}$$

and

$$q = \|T^{-1}(X^T \otimes X)\|_2, \tag{3.34}$$

where Π is the vec-permutation matrix [9, pp. 32–34]. Note that in the real case

$$p = \|T^{-1}[I_n \otimes X + (X^T \otimes I_n)\Pi]\|_2. \tag{3.35}$$

See the Appendix for a proof of the formulae (3.32)–(3.35).

4. Perturbation results for the DARE.

4.1. On perturbation equation. Let X be the unique Hermitian p.s.d. solution to the DARE (1.5), and let $\tilde{X}$ be a Hermitian solution to the perturbed DARE (1.6). Define

$$\Delta X = \tilde{X} - X, \quad \Delta Q = \tilde{Q} - Q, \quad \Delta A = \tilde{A} - A, \quad \Delta G = \tilde{G} - G.$$

Then from (1.5) and (1.6) we see that the matrix ΔX satisfies the equation

$$\Delta X - \tilde{A}^H(X + \Delta X)[I + \tilde{G}(X + \Delta X)]^{-1}\tilde{A} + A^H X(I + GX)^{-1}A - \Delta Q = 0. \tag{4.1}$$

As Konstantinov, Petkov, and Christov [15] pointed out, the higher-order term of ΔX in the perturbation equation (4.1) is hard to manipulate. In this section we shall transform the equation to an equivalent form which is easy to handle. In the course of the transformation, the matrix relations

$$(I + U)^{-1} = I - U(I + U)^{-1}, \qquad V(I + UV)^{-1} = (I + VU)^{-1}V$$

are used again and again.

Matrix operations give

$$(X + \Delta X)[I + \tilde{G}(X + \Delta X)]^{-1} = X(I + \tilde{G}X)^{-1} + (I + X\tilde{G})^{-1}\Delta X(I + \tilde{G}X)^{-1}$$

$$-(I + X\tilde{G})^{-1}\Delta X(I + \tilde{G}X)^{-1}\tilde{G}\Delta X[I + \tilde{G}(X + \Delta X)]^{-1}.$$

Consequently, (4.1) can be written as

$$\begin{aligned} &\Delta X - \tilde{A}^H(I + X\tilde{G})^{-1}\Delta X(I + \tilde{G}X)^{-1}\tilde{A} \\ &= \Delta Q + \tilde{A}^H X(I + \tilde{G}X)^{-1}\tilde{A} - A^H X(I + GX)^{-1}A \\ &\quad -\tilde{A}^H(I + X\tilde{G})^{-1}\Delta X(I + \tilde{G}X)^{-1}\tilde{G}\Delta X[I + \tilde{G}(X + \Delta X)]^{-1}\tilde{A}. \end{aligned} \tag{4.2}$$

We now define the matrices F, Φ, Ψ, K, Θ by

$$F = (I + GX)^{-1}, \qquad \Phi = FA, \quad \Psi = XF, \quad K = \Psi A, \quad \Theta = F(I + \Delta G\Psi)^{-1}. \tag{4.3}$$

Substituting $\tilde{A} = A + \Delta A, \tilde{G} = G + \Delta G$, and $\tilde{X} = X + \Delta X$ into (4.2) we see that the matrix $\Delta X \in \mathcal{H}^{n\times n}$ satisfies the equation

$$\Delta X - \tilde{\Phi}^H \Delta X \tilde{\Phi} = E + h_2(\Delta X), \tag{4.4}$$

where the matrix $\tilde{\Phi}$ is expressed by

$$\tilde{\Phi} = (I + \tilde{G}X)^{-1}\tilde{A} = (I + GX + \Delta GX)^{-1}(A + \Delta A) = \Phi + \Delta\Phi \tag{4.5}$$

with

$$\Delta\Phi = F[\Delta A - \Delta G\Psi(I + \Delta G\Psi)^{-1}(A + \Delta A)] = F(I + \Delta G\Psi)^{-1}(\Delta A - \Delta GK), \tag{4.6}$$

the matrix E is expressed by

$$E = \Delta Q + \tilde{A}^H X(I + \tilde{G}X)^{-1}\tilde{A} - A^H X(I + GX)^{-1}A = E_1 + E_2 \tag{4.7}$$

with

$$E_1 = \Delta Q + K^H \Delta A + \Delta A^H K - K^H \Delta GK \tag{4.8}$$

and

$$\begin{aligned} E_2 \quad &= \Delta A^H \Psi \Delta A + K^H \Delta G\Psi(I + \Delta G\Psi)^{-1}\Delta GK - K^H \Delta G\Psi(I + \Delta G\Psi)^{-1}\Delta A \\ &\quad -\Delta A^H \Psi(I + \Delta G\Psi)^{-1}(\Delta GK + \Delta G\Psi\Delta A), \end{aligned} \tag{4.9}$$

and the function $h_2(\Delta X)$ is expressed by

$$h_2(\Delta X) = -(A + \Delta A)^H \Theta^H \Delta X\Theta(G + \Delta G)\Delta X\Theta[I + (G + \Delta G)\Delta X\Theta]^{-1}(A + \Delta A). \tag{4.10}$$

Further, substituting (4.5)–(4.7) into (4.4) and letting

$$h_1(\Delta X) = \Delta\Phi^H \Delta X\Phi + \Phi^H \Delta X\Delta\Phi + \Delta\Phi^H \Delta X\Delta\Phi, \tag{4.11}$$

(4.4) becomes

$$\Delta X - \Phi^H \Delta X\Phi = E_1 + E_2 + h_1(\Delta X) + h_2(\Delta X). \tag{4.12}$$

Define the linear operator $\mathbf{L}$: $\mathcal{H}^{n\times n} \to \mathcal{H}^{n\times n}$ by

$$\mathbf{L}W = W - \Phi^H W\Phi, \quad W \in \mathcal{H}^{n\times n}. \tag{4.13}$$

Since the matrix Φ defined by (4.3) is d-stable, the operator $\mathbf{L}$ is invertible. Define the function $\mu(\Delta X)$ by

$$\mu(\Delta X) = \mathbf{L}^{-1}E_1 + \mathbf{L}^{-1}E_2 + \mathbf{L}^{-1}[h_1(\Delta X) + h_2(\Delta X)]. \tag{4.14}$$

Obviously, $\mu(\Delta X)$ can be regarded as a continuous mapping $\mathcal{M} : \mathcal{H}^{n\times n} \to \mathcal{H}^{n\times n}$, and the set of the solutions to (4.12) is just the set of the fixed points of the mapping $\mathcal{M}$.

Moreover, define the linear operators $\mathbf{P}$: $\mathcal{C}^{n\times n} \to \mathcal{H}^{n\times n}$, and $\mathbf{Q}$: $\mathcal{H}^{n\times n} \to \mathcal{H}^{n\times n}$ by [3], [15]

$$\mathbf{P}N = \mathbf{L}^{-1}(K^H N + N^H K), \quad N \in \mathcal{C}^{n\times n}, \tag{4.15}$$

and

$$\mathbf{Q}M = \mathbf{L}^{-1}(K^H MK), \quad M \in \mathcal{H}^{n\times n}. \tag{4.16}$$

Then by (4.8)

$$\mathbf{L}^{-1}E_1 = \mathbf{L}^{-1}\Delta Q + \mathbf{P}\Delta A - \mathbf{Q}\Delta G. \tag{4.17}$$

In the following subsections we shall derive an upper bound for some fixed points ΔX of the continuous mapping $\mathcal{M}$ expressed by (4.14) under some assumptions on $\Delta Q, \Delta A$, and ΔG, where $\Delta X = \tilde{X} - X$, in which X is the unique Hermitian p.s.d. solution to the DARE (1.5), and $\tilde{X}$ is the unique Hermitian p.s.d. solution to the perturbed DARE (1.6). For simplicity, we assume that $\tilde{Q}, \tilde{G} \geq 0$.

4.2. Estimates of some fixed points of $\mathcal{M}$. Let F, Φ, Ψ, K, Θ be the matrices defined by (4.3), and let $\mathbf{L}$, $\mathbf{P}$, and $\mathbf{Q}$ be the linear operators defined by (4.13), (4.15), and (4.16), respectively. Define

$$\begin{aligned} &l = \|\mathbf{L}^{-1}\|^{-1}, \quad p = \|\mathbf{P}\|, \quad q = \|\mathbf{Q}\|, \quad \phi = \|\Phi\|_2, \quad \psi = \|\Psi\|_2, \\ &\alpha = \|A\|_2, \quad \kappa = \|K\|_2, \quad f = \|F\|_2, \quad g = \|G\|_2. \end{aligned} \tag{4.18}$$

Moreover, we assume that ΔG satisfies

$$1 - \psi\|\Delta G\|_2 > 0, \tag{4.19}$$

and define

$$\delta = \frac{\|\Delta A\| + \kappa\|\Delta G\|}{1 - \psi\|\Delta G\|_2}. \tag{4.20}$$

Observe the following facts:

1. By (4.9) and (4.18)–(4.20),

$$\begin{aligned} \|E_2\| \quad &\leq \psi\|\Delta A\|_2\|\Delta A\| + \frac{\kappa^2\psi\|\Delta G\|_2\|\Delta G\|}{1 - \psi\|\Delta G\|_2} \\ &\quad + \frac{\kappa\psi\|\Delta A\|_2\|\Delta G\|}{1 - \psi\|\Delta G\|_2} + \frac{\psi(\kappa + \psi\|\Delta A\|_2)\|\Delta G\|_2\|\Delta A\|}{1 - \psi\|\Delta G\|_2} \\ &= \psi\delta(\|\Delta A\| + \kappa\|\Delta G\|) \equiv \epsilon_2. \end{aligned} \tag{4.21}$$

2. By (4.11), (4.6), and (4.18)–(4.20),

$$\|h_1(\Delta X)\| \leq (2\phi\|\Delta\Phi\|_2 + \|\Delta\Phi\|_2^2)\|\Delta X\| \leq \eta\|\Delta X\|, \tag{4.22}$$

where

$$\eta = f\delta(2\phi + f\delta). \tag{4.23}$$

3. Define

$$\hat{\alpha} = \frac{f(\alpha + \|\Delta A\|_2)}{1 - \psi\|\Delta G\|_2}, \quad \hat{g} = \frac{f(g + \|\Delta G\|_2)}{1 - \psi\|\Delta G\|_2}, \tag{4.24}$$

and assume that $1 - \hat{g}\|\Delta X\| > 0$. Then by (4.10), (4.3), (4.18), and (4.24)

$$\|h_2(\Delta X)\| \leq \frac{\|\Theta\|_2^3(\alpha + \|\Delta A\|_2)^2(g + \|\Delta G\|_2)\|\Delta X\|^2}{1 - (g + \|\Delta G\|_2)\|\Theta\|_2\|\Delta X\|} \leq \frac{\hat{\alpha}^2\hat{g}\|\Delta X\|^2}{1 - \hat{g}\|\Delta X\|}, \tag{4.25}$$

where it is assumed that

$$1 - \hat{g}\|\Delta X\| > 0. \tag{4.26}$$

Hence, by (4.14), (4.17), (4.21)–(4.23), and (4.25), ΔX satisfies

$$\|\Delta X\| \le \epsilon + \frac{1}{l}\left(\eta\|\Delta X\| + \frac{\hat{\alpha}^2\hat{g}\|\Delta X\|^2}{1 - \hat{g}\|\Delta X\|}\right), \tag{4.27}$$

where η and $\hat{\alpha}, \hat{g}$ are defined by (4.23) and (4.24), respectively, and ϵ is defined by

$$\epsilon = \epsilon_1 + \epsilon_2/l,$$

in which ϵ_2 is defined by (4.21), and ϵ_1 is defined by

$$\epsilon_1 = \|\Delta Q\|/l + p\|\Delta A\| + q\|\Delta G\|.$$

Let

$$\xi_* = \frac{2l\epsilon}{l - \eta + l\hat{g}\epsilon + \sqrt{(l - \eta + l\hat{g}\epsilon)^2 - 4l\hat{g}(l - \eta + \hat{\alpha}^2)\epsilon}}, \tag{4.28}$$

and

$$\mathcal{S}_{\xi_*} = \{\Delta X \in \mathcal{H}^{n\times n} \ : \ \|\Delta X\| \le \xi_*\}.$$

By the Schauder fixed-point theorem, and using the same technique described in section 3.2, we can prove that if ϵ satisfies

$$\epsilon \le \frac{(l - \eta)^2}{l\hat{g}\left(l - \eta + 2\hat{\alpha} + \sqrt{(l - \eta + 2\hat{\alpha})^2 - (l - \eta)^2}\right)}, \tag{4.29}$$

then the mapping $\mathcal{M}$ expressed by (4.14) has a fixed point $\Delta X_* \in \mathcal{S}_{\xi_*}$, i.e.,

$$\|\Delta X_*\| \le \xi_*. \tag{4.30}$$

Note that if the scalar η defined by (4.23) satisfies

$$l - \eta > 0, \tag{4.31}$$

then any $\Delta X \in \mathcal{S}_{\xi_*}$ satisfies the condition (4.26). In fact, for any $\Delta X \in \mathcal{S}_{\xi_*}$ we have

$$\begin{aligned} 1 - \hat{g}\|\Delta X\| &\ge 1 - \hat{g}\xi_* \ge 1 - \frac{2l\hat{g}\epsilon}{l - \eta + l\hat{g}\epsilon} \quad \text{(by (4.28))} \\ &= \frac{l - \eta - l\hat{g}\epsilon}{l - \eta + l\hat{g}\epsilon} \ge \frac{l - \eta - (l - \eta)^2/(l - \eta + 2\hat{\alpha})}{l - \eta + l\hat{g}\epsilon} \quad \text{(by (4.29))} \\ &= \frac{2(l - \eta)\hat{\alpha}}{(l - \eta + l\hat{g}\epsilon)(l - \eta + 2\hat{\alpha})} > 0 \quad \text{(by (4.31)).} \end{aligned}$$

4.3. The matrix $X + \Delta X_*$. Let $\Delta X_* \in \mathcal{S}_{\xi_*}$ be a fixed point of the mapping $\mathcal{M}$ expressed by (4.14). Define $Y = X + \Delta X_*$. Then from section 4.1 we see that the Hermitian matrix Y satisfies

$$Y - \tilde{A}^H Y(I + \tilde{G}Y)^{-1}\tilde{A} - \tilde{Q} = 0, \tag{4.32}$$

i.e., Y is a Hermitian solution to the DARE (1.6).

From

$$[(I + \tilde{G}Y)^{-1}\tilde{A}]^H Y(I + \tilde{G}Y)^{-1}\tilde{A} = \tilde{A}^H(I + Y\tilde{G})^{-1}Y[I - \tilde{G}Y(I + \tilde{G}Y)^{-1}]\tilde{A}$$

$$= \tilde{A}^H Y(I + \tilde{G}Y)^{-1}\tilde{A} - [Y(I + \tilde{G}Y)^{-1}\tilde{A}]^H \tilde{G}Y(I + \tilde{G}Y)^{-1}\tilde{A}$$

it follows that the relation (4.32) can be written as

$$Y - [(I+\tilde{G}Y)^{-1}\tilde{A}]^H Y(I+\tilde{G}Y)^{-1}\tilde{A} = \tilde{Q} + [Y(I+\tilde{G}Y)^{-1}\tilde{A}]^H \tilde{G}Y(I+\tilde{G}Y)^{-1}\tilde{A}, \tag{4.33}$$

where the matrix on the right-hand side is obviously a Hermitian p.s.d. matrix. Observe that

$$(I + \tilde{G}Y)^{-1}\tilde{A} = [I + (G + \Delta G)(X + \Delta X_*)]^{-1}(A + \Delta A) = \Phi + \Phi_1, \tag{4.34}$$

where $\Phi = (I + GX)^{-1}A$ is d-stable, Φ_1 can be expressed by

$$\Phi_1 = F[\Delta A - \Omega(I + \Omega)^{-1}(A + \Delta A)]$$

with

$$\Omega = \Delta G\Psi + G\Delta X_* F + \Delta G\Delta X_* F,$$

and a simple operation gives

$$\Phi_1 = F(I + \Delta G\Psi + G\Delta X_* F + \Delta G\Delta X_* F)^{-1}(\Delta A - \Delta GK - G\Delta X_*\Phi - \Delta G\Delta X_*\Phi),$$

and

$$\begin{aligned} \|\Phi_1\|_2 &\le \frac{f[\|\Delta A\| + \kappa\|\Delta G\| + \phi(g + \|\Delta G\|_2)\xi_*]}{1 - [\psi\|\Delta G\|_2 + f(g + \|\Delta G\|_2)\xi_*]} \\ &= \frac{f\delta + \phi\hat{g}\xi_*}{1 - \hat{g}\xi_*} \quad \text{(by (4.20) and (4.24))}, \end{aligned} \tag{4.35}$$

where it is assumed that

$$1 - \hat{g}\xi_* > 0. \tag{4.36}$$

Hence, by (4.34), (4.35), and Lemma 2.2, if

$$\frac{f\delta + \phi\hat{g}\xi_*}{1 - \hat{g}\xi_*} < \frac{l}{\phi + \sqrt{\phi^2 + l}}, \tag{4.37}$$

then the matrix $(I + \tilde{G}Y)^{-1}\tilde{A}$ is d-stable. In this case, by Lemma 2.6, the matrix Y, as a d-stabilizing solution to the DARE (4.32), is unique. Moreover, the Hermitian matrix Y, as a solution to (4.33), is p.s.d. [6, Prop. 2.1].

Thus, we have proved that under the conditions (4.19), (4.29), (4.31), (4.36), and (4.37), there is a unique Hermitian p.s.d. solution $\tilde{X} = Y$ to the DARE (1.6), and $\|\tilde{X} - X\| \le \xi_*$, where X is the unique Hermitian p.s.d. solution to the DARE (1.5).

4.4. The conditions (4.31) and (4.37). Note that the condition (4.31) can be deduced from the condition (4.37). In fact, from the inequality (4.37)

$$f\delta < l/(\phi + \sqrt{\phi^2 + l}),$$

which implies

$$2\phi f\delta + (f\delta)^2 < l,$$

and, by using (4.23), the last relation is equivalent to (4.31).

4.5. Perturbation bounds. Overall, we have the following theorem.

THEOREM 4.1. *Let X be the unique Hermitian p.s.d. solution to the DARE* (1.5). *Define the linear operators* $\mathbf{L}$, $\mathbf{P}$, *and* $\mathbf{Q}$ *by* (4.13), (4.15), *and* (4.16), *define* l, p, q, ϕ, ψ, α, κ, f, *and* g *by* (4.18), *and define* $\hat{\alpha}$, $\hat{g}$ *by* (4.24), *respectively. Moreover, let* $\tilde{Q} = Q + \Delta Q$, $\tilde{A} = A + \Delta A$, $\tilde{G} = G + \Delta G$ *be the coefficient matrices of the perturbed DARE* (1.6), *and let*

$$\delta = (\|\Delta A\|_2 + \kappa\|\Delta G\|_2)/(1 - \psi\|\Delta G\|_2), \quad \eta = f\delta(2\phi + f\delta),$$

$$\epsilon_1 = \frac{1}{l}\|\Delta Q\| + p\|\Delta A\| + q\|\Delta G\|, \quad \epsilon = \epsilon_1 + \frac{\psi\delta}{l}(\|\Delta A\| + \kappa\|\Delta G\|), \tag{4.38}$$

and

$$\xi_* = \frac{2l\epsilon}{l - \eta + l\hat{g}\epsilon + \sqrt{(l - \eta + l\hat{g}\epsilon)^2 - 4l\hat{g}(l - \eta + \hat{\alpha}^2)\epsilon}}. \tag{4.39}$$

If $\tilde{Q}, \tilde{G} \geq 0$, *and if*

$$1 - \psi\|\Delta G\|_2 > 0, \quad 1 - \hat{g}\xi_* > 0, \quad \frac{f\delta + \phi\hat{g}\xi_*}{1 - \hat{g}\xi_*} < \frac{l}{\phi + \sqrt{\phi^2 + l}}, \tag{4.40}$$

and

$$\epsilon < \frac{(l - \eta)^2}{l\hat{g}\left(l - \eta + 2\hat{\alpha} + \sqrt{(l - \eta + 2\hat{\alpha})^2 - (l - \eta)^2}\right)}, \tag{4.41}$$

then the DARE (1.6) *has a unique Hermitian p.s.d. solution* $\tilde{X}$, *and*

$$\|\tilde{X} - X\| \leq \frac{2l\epsilon}{l - \eta + l\hat{g}\epsilon + \sqrt{(l - \eta + l\hat{g}\epsilon)^2 - 4l\hat{g}(l - \eta + \hat{\alpha}^2)\epsilon}} = \xi_*. \tag{4.42}$$

From Theorem 4.1 we get the first-order perturbation bound for the solution X:

$$\|\tilde{X} - X\| \leq \epsilon_1 + O(\|(\Delta Q, \Delta A, \Delta G)\|^2), \quad \|(\Delta Q, \Delta A, \Delta G)\| \to 0, \tag{4.43}$$

where ϵ_1 is defined by (4.38). Consequently, for sufficiently small $\|(\Delta Q, \Delta A, \Delta G)\|$, we have

$$\frac{\|\tilde{X} - X\|}{\|X\|} \lesssim \frac{\|Q\|}{l\|X\|}\frac{\|\Delta Q\|}{\|Q\|} + \frac{p\|A\|}{\|X\|}\frac{\|\Delta A\|}{\|A\|} + \frac{q\|G\|}{\|X\|}\frac{\|\Delta G\|}{\|G\|}. \tag{4.44}$$

For comparing our results with those of [15], we now cite a result of [15], which also presents a perturbation bound for the Hermitian p.s.d. solution to the DARE (1.5).

THEOREM 4.2 (see [15, Thm. 3.2]). *Let X be the unique Hermitian p.s.d. solution to the DARE* (1.5). $\|\ \|$ *stands for the Frobenius norm, or the spectral norm. Define the linear operator* $\mathbf{L}$ *by* (4.13)*, and let* $l = \|\mathbf{L}^{-1}\|^{-1}$. *Moreover, let* $\tilde{Q} = Q + \Delta Q$, $\tilde{A} = A + \Delta A$, *and* $\tilde{G} = G + \Delta G$ *be the coefficient matrices of the perturbed DARE* (1.6)*, and define* a_0, a_1, a_2 *by*

$$a_0 = \left(\|\Delta Q\| + \|X\|(2\|A\| + \|\Delta A\|)\|\Delta A\| + \|A\|^2\|X\|^2\|\Delta G\|\right)/l,$$

$$a_1 = \left((2\|A\| + \|\Delta A\|)\|\Delta A\| + 2\|A\|^2\|X\|\|\Delta G\|\right)/l,$$

$$a_2 = \left(\|(I + GX)^{-1}A\|^2\|G\| + \|A\|^2\|\Delta G\|\right)/l.$$

If

$$D \equiv (1 - a_1)^2 - 4a_0a_2 > 0,$$

then there is a unique Hermitian solution $\tilde{X}$ *to the DARE* (1.6) *such that*

$$\|\tilde{X} - X\| \le \frac{1 - a_1 - \sqrt{D}}{2a_2} \equiv \xi_{\mathrm{KPC}}. \tag{4.45}$$

Remark 4.3. From Theorem 4.1 it follows that for sufficiently small $\Delta Q, \Delta A, \Delta G$, and for any unitarily invariant norm $\|\ \|$, we have the estimate

$$\|\tilde{X} - X\| \lesssim \frac{1}{l}\|\Delta Q\| + p\|\Delta A\| + q\|\Delta G\| \equiv \epsilon_1, \tag{4.46}$$

where $\tilde{X}$ is the unique Hermitian p.s.d. solution to the DARE (1.6). However, by Theorem 4.2, the matrix $\tilde{X}$ of (4.45) is only a Hermitian solution to the perturbed DARE (1.6). It is not considered whether the solution $\tilde{X}$ is p.s.d. Moreover, from Theorem 4.2 it follows that for sufficiently small ΔQ, ΔA, ΔG, we have the estimate [15, eq. (37)]

$$\|\tilde{X} - X\| \lesssim \frac{1}{l}(\|\Delta Q\| + 2\|A\|\|X\|\|\Delta A\| + \|A\|^2\|X\|^2\|\Delta G\|) \equiv \epsilon_{\mathrm{KPC}}, \tag{4.47}$$

where $\|\ \|$ stands for the Frobenius norm, or the spectral norm. Observe that from (4.15), (4.16), (4.3), and (4.18)

$$\begin{aligned} &p \le 2\|K\|/l \le 2\|X(I + GX)^{-1}\|\|A\|/l \le 2\|X\|\|A\|/l, \\ &q \le \|K\|^2/l \le \|X(I + GX)^{-1}\|^2\|A\|^2/l \le \|X\|^2\|A\|^2/l, \end{aligned} \tag{4.48}$$

where the last inequalities of (4.48) hold is due to the fact that for the Hermitian p.s.d. matrices X and G, and for any positive scalar μ, we have

$$(X + \mu I)[I + G(X + \mu I)]^{-1} = [(X + \mu I)^{-1} + G]^{-1} \le X + \mu I,$$

and

$$\|(X + \mu I)[I + G(X + \mu I)]^{-1}\| \le \|X + \mu I\|. \tag{4.49}$$

Taking $\mu \to 0$, from (4.49) we get $\|X(I+GX)^{-1}\| \le \|X\|$. Hence, the estimate (4.46) implies (4.47) (by (4.48)). Consequently, the result of Theorem 4.1 is better than that of Theorem 4.2.

4.6. Condition numbers. The relation (4.44) shows that the scalars $c_Q(X)$, $c_A(X)$, $c_G(X)$, and $c_Q^{(r)}(X)$, $c_A^{(r)}(X)$, $c_G^{(r)}(X)$, defined by

$$c_Q(X) = \frac{1}{l}, \quad c_A(X) = p, \quad c_G(X) = q$$

and

$$c_Q^{(r)}(X) = \frac{\|Q\|}{l\|X\|}, \quad c_A^{(r)}(X) = \frac{p\|A\|}{\|X\|}, \quad c_G^{(r)}(X) = \frac{q\|G\|}{\|X\|}$$

are the absolute and relative condition numbers of X with respect to Q, A, G, respectively. Further, the scalar $c^{(r)}(X)$ defined by

$$\begin{aligned} c^{(r)}(X) &= \sqrt{[c_Q^{(r)}(X)]^2 + [c_A^{(r)}(X)]^2 + [c_G^{(r)}(X)]^2} \\ &= \frac{1}{\|X\|}\sqrt{(\|Q\|/l)^2 + (p\|A\|)^2 + (q\|G\|)^2} \end{aligned} \tag{4.50}$$

can be regarded as the relative condition number of the solution X.

4.7. Expressions of $\|\mathbf{L}^{-1}\|^{-1}$, $\|\mathbf{P}\|$, and $\|\mathbf{Q}\|$. Let $\mathbf{L}$, $\mathbf{P}$, and $\mathbf{Q}$ be the linear operators defined by (4.13), (4.15), and (4.16), respectively, and let $l = \|\mathbf{L}^{-1}\|^{-1}$, $p = \|\mathbf{P}\|$, and $q = \|\mathbf{Q}\|$ (see (4.18)). The problem of finding explicit expressions of l, p, and q is a difficult one. However, if the operator norm $\|\ \|$ for defining l and q is induced by the Frobenius norm $\|\ \|_F$ on $\mathcal{H}^{n\times n}$, and the operator norm $\|\ \|$ for defining p is induced by the Frobenius norm $\|\ \|_F$ on $\mathcal{C}^{n\times n}$, then we can find explicit expressions of the corresponding l, p, and q: define the matrix T by

$$T = I_{n^2} - \Phi^T \otimes \Phi^H, \tag{4.51}$$

where $\Phi = (I + GX)^{-1}A$ is d-stable. Then

$$l = \|T^{-1}\|_2^{-1}, \tag{4.52}$$

$$p = \|T^{-1}[(I_n \otimes K^H)(I_{n^2},\, iI_{n^2}) + (K^T \otimes I_n)\Pi(I_{n^2},\, -iI_{n^2})]\|_2, \tag{4.53}$$

and

$$q = \|T^{-1}(K^T \otimes K^H)\|_2, \tag{4.54}$$

where Π is the vec-permutation matrix [9, pp. 32–34], and $K = X(I+GX)^{-1}A$. Note that in the real case

$$p = \|T^{-1}[I_n \otimes K^T + (K^T \otimes I_n)\Pi]\|_2. \tag{4.55}$$

See the Appendix for a proof of the formulae (4.52)–(4.55).

5. Numerical examples. We now use simple numerical examples to illustrate our results. All computations were performed using MATLAB, version 4.2c, implemented on a SALT. The relative machine precision reported by MATLAB is 2.2204×10^{-16}.

Example 5.1 (see [3, Example 2]). Consider the CARE (1.3) with

$$A = \begin{pmatrix} -0.100 & 0.000 \\ 0.000 & -0.020 \end{pmatrix}, \quad Q = C^T C \quad \text{with} \quad C = (10,\ 100),$$

$$G = BR^{-1}B^T \quad \text{with} \quad B = \begin{pmatrix} 0.100 & 0.000 \\ 0.001 & 0.010 \end{pmatrix}, \quad R = \begin{pmatrix} 1+10^{-m} & 1 \\ 1 & 1 \end{pmatrix}.$$

The pair (A, G) is c-stabilizable, and the pair (A, Q) is c-detectable. Suppose that the perturbations in the coefficient matrices are

$$\Delta Q = \Delta Q_0 \times 10^{-j}, \quad \Delta A = \Delta A_0 \times 10^{-j}, \quad \Delta G = \Delta G_0 \times 10^{-j}$$

with

$$\Delta Q_0 = \begin{pmatrix} 5 & -2 \\ -2 & 4 \end{pmatrix}, \quad \Delta A_0 = \begin{pmatrix} 0.3 & -0.2 \\ 0.1 & 0.1 \end{pmatrix}, \quad \Delta G_0 = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & -0.3 \end{pmatrix}.$$

Let $\tilde{Q} = Q + \Delta Q$, $\tilde{A} = A + \Delta A$, and $\tilde{G} = G + \Delta G$ be the coefficient matrices of the perturbed CARE (1.4).

By using the MATLAB file "are" one can compute the unique Hermitian p.s.d. solution X to the CARE (1.3) and the unique Hermitian p.s.d. solution $\tilde{X}$ to the perturbed CARE (1.4). Some numerical results on relative perturbation bounds $\epsilon/\|X\|_F$ and $\xi_*/\|X\|_F$ are listed in Table 5.1, where ϵ and ξ_* are as in (3.25) and (3.24) (see (3.21) and (3.22) for the definitions). The scalars $\chi^{(l)}$ and $\chi^{(u)}$ are the lower and upper bounds for χ, the relative perturbation bound given by Xu [32] (see (3.27) and (3.28)). The relative condition number $c^{(r)}(X)$ of the solution X is defined by (3.29). The scalars l, p, and q are computed by the formulae (3.32), (3.35), and (3.34), respectively. The cases when the condition (3.23) of Theorem 3.1 is violated are denoted by asterisks.

TABLE 5.1
$j = 12$.

m	$\frac{\|\tilde{X}-X\|_F}{\|X\|_F}$	$\epsilon/\|X\|_F$	$\xi_*/\|X\|_F$	$\chi^{(l)}$	$\chi^{(u)}$	$c^{(r)}(X)$
0	1.3707e-09	1.9142e-09	1.9142e-09	2.0984e-08	4.1968e-08	5.0007e+01
1	1.3830e-09	1.9305e-09	1.9306e-09	9.9505e-08	1.9901e-07	5.0254e+02
2	1.4525e-09	2.0274e-09	2.0316e-09	5.7608e-06	1.1522e-05	5.2749e+03
3	1.6101e-09	2.2534e-09	2.4817e-09	1.9129e-04	3.8259e-04	5.8481e+04
4	1.7457e-09	2.4482e-09	*	3.1086e-03	6.2171e-03	6.3406e+05
5	1.7962e-09	2.5390e-09	*	3.6628e-02	7.3256e-02	6.5701e+06

The results listed in Table 5.1 show that the relative perturbation bounds $\epsilon/\|X\|_F$ and $\xi_*/\|X\|_F$ are fairly sharp, and the bound χ given by [32] is conservative.

Example 5.2 (see [15]). Consider the DARE (1.5) with

$$Q = VQ_0V, \quad A = VA_0V, \quad G = VG_0V,$$

where

$$Q_0 = \mathrm{diag}(10^m, 1, 10^{-m}), \quad A_0 = \mathrm{diag}(0, 10^{-m}, 1), \quad G_0 = \mathrm{diag}(10^{-m}, 10^{-m}, 10^{-m}),$$

and

$$V = I - 2vv^T/3, \quad v = (1, 1, 1)^T.$$

The perturbations in the coefficient matrices are

$$\Delta Q = V\Delta Q_0 V, \quad \Delta A = V\Delta A_0 V, \quad \Delta G = V\Delta G_0 V,$$

where

$$\Delta Q_0 = \begin{pmatrix} 10^m & -5 & 7 \\ -5 & 1 & 3 \\ 7 & 3 & 10^m \end{pmatrix} \times 10^{-j}, \quad \Delta A_0 = \begin{pmatrix} 3 & -4 & 8 \\ -6 & 2 & -9 \\ 2 & 7 & 5 \end{pmatrix} \times 10^{-j},$$

and

$$\Delta G_0 = \begin{pmatrix} 10^{-m} & -10^{-m} & 2\times 10^{-m} \\ -10^{-m} & 5\times 10^{-m} & -10^{-m} \\ 2\times 10^{-m} & -10^{-m} & 3\times 10^{-m} \end{pmatrix} \times 10^{-j}.$$

The unique Hermitian p.s.d. solution X to the DARE (1.5) is given by $X = VX_0V$, where $X_0 = \mathrm{diag}(x_1, x_2, x_3)$ with

$$x_i = \{a_i^2 + q_i g_i - 1 + [(a_i^2 + q_i g_i - 1)^2 + 4q_i g_i]^{1/2}\}/(2g_i),$$

and q_i, a_i, and g_i are the corresponding diagonal elements of Q_0, A_0, and G_0.

Let $\tilde{Q} = Q + \Delta Q$, $\tilde{A} = A + \Delta A$, and $\tilde{G} = G + \Delta G$ be the coefficient matrices of the perturbed DARE (1.6). By using MATLAB and the file "dare" one can compute the unique Hermitian p.s.d. solution $\tilde{X}$ to (1.6). Note that the file "dare" is a computer program written by Alan J. Laub (1993). The program is an implementation of a generalized eigenproblem algorithm by Arnold and Laub [2].

Some numerical results on relative perturbation bounds are listed in Tables 5.2 and 5.3, where the bounds ϵ_1, ξ_*, ϵ_{KPC}, and ξ_{KPC} are defined by (4.46), (4.42), (4.47), and (4.45), respectively. The relative condition number $c^{(r)}(X)$ of the solution X is defined by (4.50). The scalars l, p, and q are computed by the formulae (4.52), (4.55), and (4.54). The cases when the conditions of Theorem 4.1 or Theorem 4.2 are violated are denoted by asterisks.

TABLE 5.2
$m = 2,\ c^{(r)}(X) \approx 47.$

j	$\frac{\|\tilde{X}-X\|_F}{\|X\|_F}$	$\epsilon_1/\|X\|_F$	$\xi_*/\|X\|_F$	$\epsilon_{\mathrm{KPC}}/\|X\|_F$	$\xi_{\mathrm{KPC}}/\|X\|_F$
10	5.5516e-09	8.2724e-09	8.2724e-09	1.9863e-07	1.9863e-07
9	5.5516e-08	8.2724e-08	8.2724e-08	1.9863e-06	1.9866e-06
8	5.5514e-07	8.2724e-07	8.2728e-07	1.9863e-05	1.9895e-05
7	5.5503e-06	8.2724e-06	8.2767e-06	1.9863e-04	2.0193e-04
6	5.5394e-05	8.2724e-05	8.3163e-05	1.9863e-03	2.4854e-03
5	5.4352e-04	8.2724e-04	8.7518e-04	1.9863e-02	*
4	4.7144e-03	8.2724e-03	*	1.9863e-01	*

The results listed in Tables 5.2 and 5.3 show that the relative perturbation bounds $\epsilon_1/\|X\|_F$ and $\xi_*/\|X\|_F$ are sharper than the bounds $\epsilon_{\mathrm{KPC}}/\|X\|_F$ and $\xi_{\mathrm{KPC}}/\|X\|_F$ given by [15].

Appendix. We provide proof of the formulae (3.32)–(3.35) and (4.52)–(4.55). We first cite a lemma of [4].

TABLE 5.3
$j = 10$.

m	$\frac{\|\tilde{X}-X\|_F}{\|X\|_F}$	$\epsilon_1/\|X\|_F$	$\xi_*/\|X\|_F$	$\frac{\epsilon_{\rm KPC}}{\|X\|_F}$	$\frac{\xi_{\rm KPC}}{\|X\|_F}$	$c^{(r)}(X)$
0	7.5367e-10	1.7284e-09	1.7284e-09	1.0326e-08	1.0326e-08	1.2038e+00
1	1.0742e-09	2.6635e-09	2.6635e-09	2.2239e-08	2.2239e-08	5.2360e+00
2	5.5516e-09	8.2724e-09	8.2724e-09	1.9863e-07	1.9863e-07	4.7054e+01
3	5.0549e-08	6.7488e-08	6.7490e-08	1.9671e-06	1.9702e-06	4.6601e+02
4	4.9931e-07	6.5999e-07	6.6204e-07	1.9652e-05	2.4487e-05	4.6557e+03
5	4.1427e-06	6.5851e-06	$*$	1.9650e-04	$*$	4.6552e+04

LEMMA A.1 (see [4, Lemma 7]). *If $W_1, W_2 \in \mathcal{H}^{n\times n}$ satisfy $W_2 \geq W_1 \geq -W_2$, then $\|W_1\|_F \leq \|W_2\|_F$.*

Let $\Phi \in \mathcal{C}^{n\times n}$ be c-stable, $\mathbf{L}$, $\mathbf{P}$, and $\mathbf{Q}$ be the linear operators defined by (3.4), (3.7), and (3.8), and let $l = \|\mathbf{L}^{-1}\|^{-1}$, $p = \|\mathbf{P}\|$, and $q = \|\mathbf{Q}\|$, where the operator norm $\|\ \|$ for defining l and q is induced by the Frobenius norm $\|\ \|_F$ on $\mathcal{H}^{n\times n}$, and the operator norm $\|\ \|$ for defining p is induced by the Frobenius norm $\|\ \|_F$ on $\mathcal{C}^{n\times n}$.

Define the linear operator $\mathbf{T}$: $\mathcal{C}^{n\times n} \to \mathcal{C}^{n\times n}$ by

$$\mathbf{T}Z = \Phi^H Z + Z\Phi, \quad Z \in \mathcal{C}^{n\times n}. \tag{A-1}$$

Byers and Nash [4, Thm. 8] prove the formula (3.32): $l = \|T^{-1}\|_2^{-1}$, where T, the matrix representation of $\mathbf{T}$, is expressed by (3.31).

By (3.7)

$$\mathbf{P}N = \mathbf{L}^{-1}(XN + N^H X) \equiv W \in \mathcal{H}^{n\times n}, \quad N \in \mathcal{C}^{n\times n}.$$

Combining it with

$$Z \equiv \mathbf{T}^{-1}(XN + N^H X)$$

follows that for the same $N \in \mathcal{C}^{n\times n}$ we have $\mathbf{T}Z = \mathbf{L}W$, or, equivalently,

$$\Phi^H(Z - W) + (Z - W)\Phi = 0. \tag{A-2}$$

Since Φ is c-stable, (A-2) implies $W = Z$, i.e.,

$$\mathbf{P}N = \mathbf{T}^{-1}(XN + N^H X), \quad N \in \mathcal{C}^{n\times n}.$$

Consequently,

$$\begin{aligned} p &= \max_{\substack{N \in \mathcal{C}^{n\times n} \\ N \neq 0}} \frac{\|\mathbf{T}^{-1}(XN + N^H X)\|_F}{\|N\|_F} \\ &= \max_{\substack{N \in \mathcal{C}^{n\times n} \\ N \neq 0}} \frac{\|T^{-1}[(I_n \otimes X)\mathrm{vec}N + (X^T \otimes I_n)\mathrm{vec}N^H]\|_2}{\|\mathrm{vec}N\|_2}. \end{aligned} \tag{A-3}$$

Write $N = N_R + iN_I$ with $N_R, N_I \in \mathcal{R}^{n\times n}$ and $i = \sqrt{-1}$. Observe that

$$\mathrm{vec}N = (I_{n^2},\ iI_{n^2}) \begin{pmatrix} \mathrm{vec}N_R \\ \mathrm{vec}N_I \end{pmatrix}, \quad \mathrm{vec}N^H = \Pi(I_{n^2},\ -iI_{n^2}) \begin{pmatrix} \mathrm{vec}N_R \\ \mathrm{vec}N_I \end{pmatrix},$$

where Π is the vec-permutation matrix. Hence, from (A-3) we get the formula (3.33). Note that in the real case, $N \in \mathcal{R}^{n\times n}$ in (A-3). Consequently, we have the formula (3.35).

Let $\mathbf{T}$ be the linear operator defined by (A-1), and define the linear operator $\mathbf{R}$: $\mathcal{C}^{n\times n} \to \mathcal{C}^{n\times n}$ by

$$\mathbf{R}N = \mathbf{T}^{-1}(XNX), \quad N \in \mathcal{C}^{n\times n},$$

where $X \in \mathcal{H}^{n\times n}$. We now prove the formula (3.34): $q = \|T^{-1}(X^T \otimes X)\|_2$.

Obviously, we only need to prove the relation

$$\max_{\substack{H \in \mathcal{H}^{n\times n}\\ H \neq 0}} \frac{\|\mathbf{L}^{-1}(XHX)\|_F}{\|H\|_F} = \max_{\substack{N \in \mathcal{C}^{n\times n}\\ N \neq 0}} \frac{\|\mathbf{T}^{-1}(XNX)\|_F}{\|N\|_F}. \tag{A-4}$$

We first prove that there exists a matrix $N_* \in \mathcal{C}^{n\times n}$ such that

$$\max_{\substack{N \in \mathcal{C}^{n\times n}\\ N \neq 0}} \frac{\|\mathbf{T}^{-1}(XNX)\|_F}{\|N\|_F} = \frac{\|\mathbf{T}^{-1}(XN_*X)\|_F}{\|N_*\|_F}, \tag{A-5}$$

and either $N_*^H = N_*$ or $N_*^H = -N_*$.

Since the operator $\mathbf{R}$ is a linear transformation on the vector space $\mathcal{C}^{n\times n}$, and the Frobenius norm is just the Euclidean vector norm applied to "vectors" in $\mathcal{C}^{n\times n}$, so the maximum in the left-hand side of (A-5) occurs when N is a singular "vector" of $\mathbf{R}$ corresponding to the largest singular value. Let $N_1 \in \mathcal{C}^{n\times n}$ be such a singular "vector," and let

$$Z_1 = \mathbf{T}^{-1}(XN_1X). \tag{A-6}$$

Then by (A-1) we can write (A-6) as

$$\Phi^H Z_1 + Z_1\Phi = XN_1X,$$

or, equivalently,

$$\Phi^H Z_1^H + Z_1^H\Phi = XN_1^HX, \quad \text{i.e.,} \quad Z_1^H = \mathbf{T}^{-1}(XN_1^HX).$$

Thus, we have

$$\|\mathbf{T}^{-1}(XN_1^HX)\|_F = \|\mathbf{T}^{-1}(XN_1X)\|_F.$$

This means that N_1^H is also a singular "vector" of $\mathbf{R}$ corresponding to the largest singular value. If $N_1 + N_1^H = 0$, then $N_* = N_1$ is a skew-Hermitian matrix satisfying (A-5). Otherwise, $N_* = N_1 + N_1^H$ is a Hermitian matrix satisfying (A-5).

Therefore, for proving (A-4) we only need to show that for any skew-Hermitian $K \in \mathcal{C}^{n\times n}$, there is a matrix $H \in \mathcal{H}^{n\times n}$ such that

$$\frac{\|\mathbf{T}^{-1}(XKX)\|_F}{\|K\|_F} \leq \frac{\|\mathbf{L}^{-1}(XHX)\|_F}{\|H\|_F}. \tag{A-7}$$

Let

$$Z_1 = \mathbf{T}^{-1}(XKX), \tag{A-8}$$

and let

$$W_1 = iZ_1, \quad \Omega_1 = X(iK)X. \tag{A-9}$$

Then by (A-1) and (A-9), the relation (A-8) can be written as

$$\Phi^H W_1 + W_1\Phi = \Omega_1. \tag{A-10}$$

Decompose the Hermitian matrix iK as

$$iK = U\mathrm{diag}(\lambda_1, \ldots, \lambda_n)U^H, \tag{A-11}$$

where $U \in \mathcal{C}^{n\times n}$ is unitary, and λ_j are real scalars. Further, define $H \in \mathcal{H}^{n\times n}$ by

$$H = U\mathrm{diag}(|\lambda_1|, \ldots, |\lambda_n|)U^H, \tag{A-12}$$

and define $\Omega_2 \in \mathcal{H}^{n\times n}$ by

$$\Omega_2 = XHX \geq 0. \tag{A-13}$$

Then from (A-11) and (A-12)

$$H \geq iK \geq -H, \quad \|H\|_F = \|K\|_F, \tag{A-14}$$

and from (A-9), (A-13), and (A-14)

$$\Omega_2 \geq \Omega_1 \geq -\Omega_2. \tag{A-15}$$

If W_2 solves

$$\Phi^H W_2 + W_2\Phi = -\Omega_2, \tag{A-16}$$

then W_2 is Hermitian p.s.d. [31, Lem. 12.1], and by the definition of $\mathbf{L}$, W_2 can be expressed by

$$W_2 = -\mathbf{L}^{-1}\Omega_2. \tag{A-17}$$

Combining (A-16) with (A-10) gives

$$\Phi^H(W_2 - W_1) + (W_2 - W_1)\Phi = -(\Omega_2 + \Omega_1), \tag{A-18}$$

and

$$\Phi^H(W_2 + W_1) + (W_2 + W_1)\Phi = -(\Omega_2 - \Omega_1), \tag{A-19}$$

where $\Omega_2 + \Omega_1 \geq 0$, and $\Omega_2 - \Omega_1 \geq 0$ (by (A-15)). Since Φ is c-stable, (A-18) and (A-19) imply that both the matrices $W_2 + W_1$ and $W_2 - W_1$ are Hermitian p.s.d. [31, Lem. 12.1], i.e.,

$$W_2 \geq W_1 \geq -W_2.$$

By Lemma A.1

$$\|W_1\|_F \leq \|W_2\|_F. \tag{A-20}$$

Consequently, we have

$$\begin{aligned}
\frac{\|\mathbf{T}^{-1}(XKX)\|_F}{\|K\|_F} &= \frac{\|Z_1\|_F}{\|K\|_F} \quad \text{(by (A-8))} \\
&= \frac{\|W_1\|_F}{\|H\|_F} \quad \text{(by (A-9) and (A-14))} \\
&\le \frac{\|W_2\|_F}{\|H\|_F} \quad \text{(by (A-20))} \\
&= \frac{\|\mathbf{L}^{-1}\Omega_2\|_F}{\|H\|_F} \quad \text{(by (A-17))} \\
&= \frac{\|\mathbf{L}^{-1}(XHX)\|_F}{\|H\|} \quad \text{(by (A-13))},
\end{aligned}$$

where $H \in \mathcal{H}^{n\times n}$ is defined by (A-12). Thus, (A-7) is proved.

Let $\Phi \in \mathcal{C}^{n\times n}$ be d-stable, $\mathbf{L}$, $\mathbf{P}$, and $\mathbf{Q}$ be the linear operators defined by (4.13), (4.15), and (4.16), and let $l = \|\mathbf{L}^{-1}\|^{-1}$, $p = \|\mathbf{P}\|$, and $q = \|\mathbf{Q}\|$, where the operator norm $\|\ \|$ for defining l and q is induced by the Frobenius norm $\|\ \|_F$ on $\mathcal{H}^{n\times n}$, and the operator norm $\|\ \|$ for defining p is induced by the Frobenius norm $\|\ \|_F$ on $\mathcal{C}^{n\times n}$. Moreover, define the linear operator $\mathbf{T}$: $\mathcal{C}^{n\times n} \to \mathcal{C}^{n\times n}$ by

$$\mathbf{T}Z = Z - \Phi^H Z\Phi, \quad Z \in \mathcal{C}^{n\times n}. \tag{A-21}$$

The formulae (4.52)–(4.55) can be proved by using the same technique described above. For example, we now prove the formula (4.52): $l = \|T^{-1}\|_2^{-1}$, where T, the matrix representation of $\mathbf{T}$, is expressed by (4.51).

Obviously, we only need to prove the relation

$$\min_{\substack{Z\in\mathcal{C}^{n\times n}\\ Z\neq 0}} \frac{\|Z - \Phi^H Z\Phi\|_F}{\|Z\|} = \min_{\substack{W\in\mathcal{H}^{n\times n}\\ W\neq 0}} \frac{\|W - \Phi^H W\Phi\|_F}{\|W\|}. \tag{A-22}$$

First of all, by using the technique described by [4, proof of Lem. 1] we can prove that there exists a matrix $Z_* \in \mathcal{C}^{n\times n}$ such that

$$\min_{\substack{Z\in\mathcal{C}^{n\times n}\\ Z\neq 0}} \frac{\|Z - \Phi^H Z\Phi\|_F}{\|Z\|} = \frac{\|Z_* - \Phi^H Z_*\Phi\|_F}{\|Z_*\|},$$

and either $Z_*^H = Z_*$ or $Z_*^H = -Z_*$. Therefore, for proving (A-22) we only need to show that for any skew-Hermitian $K \in \mathcal{C}^{n\times n}$, there is a matrix $H \in \mathcal{H}^{n\times n}$ such that

$$\frac{\|K - \Phi^H K\Phi\|_F}{\|K\|} \ge \frac{\|H - \Phi^H H\Phi\|_F}{\|H\|},$$

which can be proved by a similar argument as above for (A-7).

Acknowledgments. I am grateful to the referee for very helpful comments and suggestions. I am also grateful to Volker Mehrmann and Peter Benner for a solver for the DARE.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice–Hall, Englewood Cliffs, NJ, 1979.

[2] W. F. Arnold III and A. J. Laub, *Generalized eigenproblem algorithms and software for algebraic Riccati equations*, Proc. IEEE, 72 (1984), pp. 1746–1754.

[3] R. Byers, *Numerical condition of the algebraic Riccati equation*, in Linear Algebra and Its Role in System Theory, B. N. Datta, ed., Contemp. Math. 47, AMS, Providence, RI, 1985, pp. 35–49.

[4] R. Byers and S. Nash, *On the singular "vectors" of the Lyapunov operator*, SIAM J. Alg. Disc. Meth., 8 (1987), pp. 59–66.

[5] C.-H. Chen, *Perturbation analysis for solutions of algebraic Riccati equations*, J. Comput. Math., 6 (1988), pp. 336–347.

[6] P. M. Gahinet, A. J. Laub, C. S. Kenney, and G. A. Hewer, *Sensitivity of the stable discrete-time Lyapunov equation*, IEEE Trans. Automat. Control, 35 (1990), pp. 1209–1217.

[7] A. R. Ghavimi and A. J. Laub, *Backward error, sensitivity, and refinement of computed solutions of algebraic Riccati equations*, Numer. Linear Algebra Appl., 2 (1995), pp. 29–49.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[9] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, John Wiley, New York, 1981.

[10] T. Gudmundsson, C. Kenney, and A. J. Laub, *Scaling of the discrete-time algebraic Riccati equation to enhance stability of the Schur solution method*, IEEE Trans. Automat. Control, 37 (1992), pp. 513–518.

[11] V. Ionescu and M. Weiss, *On computing the stabilizing solution of the discrete-time Riccati equation*, Linear Algebra Appl., 174 (1992), pp. 229–238.

[12] C. Kenney and G. Hewer, *The sensitivity of the algebraic and differential Riccati equations*, SIAM J. Control Optim., 28 (1990), pp. 50–69.

[13] C. Kenney, A. J. Laub and M. Wette, *Error bounds for Newton refinement of solutions to algebraic Riccati equations*, Math. Control Signals Systems, 3 (1990), pp. 211–224.

[14] M. M. Konstantinov, P. H. Petkov, and N. D. Christov, *Perturbation analysis of matrix quadratic equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 1159–1163.

[15] M. M. Konstantinov, P. H. Petkov, and N. D. Christov, *Perturbation analysis of the discrete Riccati equation*, Kybernetika, 29 (1993), pp. 18–29.

[16] M. M. Konstantinov, P. H. Petkov, D. W. Gu, and I. Postlethwaite, *Perturbation Techniques for Linear Control Problems*, Report 95-7, Control Systems Research, Department of Engineering, Leicester University, UK, 1995.

[17] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, New York, 1995.

[18] A. J. Laub, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 913–921.

[19] V. Mehrmann, *A symplectic orthogonal method for single input or single output discrete time optimal quadratic control problems*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 221–247.

[20] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[21] T. Pappas, A. J. Laub, and N. R. Sandell, Jr., *On the numerical solution of the discrete-time algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 631–641.

[22] R. V. Patel, A. J. Laub, and P. M. Van Dooren, *Introduction and survey*, in Numerical Linear Algebra Techniques for Systems and Control, A Selected Reprint Volume, IEEE Control Systems Society, New York, 1994.

[23] A. P. Sage and C. C. White, *Optimum System Control*, Prentice–Hall, Englewood Cliffs, NJ, 1977.

[24] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[25] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[26] J.-G. Sun, *Residual bounds of approximate solution of the algebraic Riccati equation*, Numer. Math., 76 (1997), pp. 249–263.

[27] J.-G. Sun, *Residual bounds of approximate solutions of the discrete-time algebraic Riccati equation*, Numer. Math., to appear.

[28] J.-G. Sun, *Backward error for the discrete-time algebraic Riccati equation*, Linear Algebra

Appl., 259 (1997), pp. 183–208.
[29] P. Van Dooren, *A generalized eigenvalue approach for solving Riccati equations*, SIAM J. Sci. Statist. Comput., 2 (1981), pp. 121–135.
[30] C. Van Loan, *How near is a stable matrix to an unstable matrix?*, Contemp. Math., 47 (1985), pp. 465–478.
[31] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979.
[32] S.-F. Xu, *Sensitivity analysis of the algebraic Riccati equations*, Numer. Math., 75 (1996), pp. 121–134.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 66–70, January 1998

004

INEQUALITIES FOR THE HADAMARD PRODUCT OF MATRICES*

B. MOND† AND J. E. PEČARIĆ‡

Abstract. Some known inequalities for the Hadamard product of matrices are extended and new inequalities obtained.

Key words. matrix inequalities, Hadamard products

AMS subject classification. 15A45

PII. S0895479896302953

1. Introduction. If A, B are positive semidefinite $n \times n$ Hermitian matrices, then $A^2 \circ B^2 - (A \circ B)^2$ is positive semidefinite, i.e., in inequality form, we have [1]

$$(A \circ B)^2 \leq A^2 \circ B^2, \tag{1.1}$$

where $A \circ B$ is the Hadamard product of matrices A and B.

From this inequality, we can also get [1]

$$A \circ B \leq (A^2 \circ B^2)^{1/2} \tag{1.2}$$

and

$$A^{1/2} \circ B^{1/2} \leq (A \circ B)^{1/2}. \tag{1.3}$$

Some converse results were obtained recently in [4]. We have

$$A^2 \circ B^2 - (A \circ B)^2 \leq \frac{1}{4}(M - m)^2 I, \tag{1.4}$$

and

$$(A^2 \circ B^2)^{1/2} \leq \frac{M + m}{2\sqrt{Mm}} A \circ B, \tag{1.5}$$

where A and B are positive definite Hermitian matrices, and M and m are, respectively, the largest and smallest eigenvalues of $A \otimes B$ (the Kronecker product of A and B).

Some generalizations and related results will be given in this paper.

2. Results.

THEOREM 2.1. *Let A and B be positive definite $n \times n$ Hermitian matrices and let r and s be two nonzero integers such that $s > r$. Then*

$$(A^s \circ B^s)^{1/s} \geq (A^r \circ B^r)^{1/r}. \tag{2.1}$$

*Received by the editors May 3, 1996; accepted for publication (in revised form) by P. Lancaster December 18, 1996.

http://www.siam.org/journals/simax/19-1/30295.html

†Department of Mathematics, La Trobe University, Bundoora, Victoria, 3083, Australia (b.mond@latrobe.edu.au).

‡Faculty of Textile Technology, University of Zagreb, Zagreb, Croatia (pecaric@mahazu.hazu.hr).

Proof. The following result holds [5].

Let A be an $n \times n$ positive definite Hermitian matrix and let V be an $n \times t$ matrix such that $V^*V = I$. Then

$$(V^*A^sV)^{1/s} \geq (V^*A^rV)^{1/r} \tag{2.2}$$

for all real r and s such that $s \notin (-1, 1)$ and $r \notin (-1, 1),\ s > r$.

In our case, nonzero integers r and s satisfy these conditions. Further, instead of V, we use J, the selection matrix of order $n^2 \times n$ with the property [3, 4]

$$A \circ B = J^t(A \otimes B)J \tag{2.3}$$

as well as the fact that for any integer p we have

$$(A \otimes B)^p = A^p \otimes B^p. \tag{2.4}$$

Thus (2.2) gives

$$(J^t(A \otimes B)^sJ)^{1/s} \geq (J^t(A \otimes B)^r)^{1/r}$$

and, from (2.4),

$$(J^t(A^s \otimes B^s)J)^{1/s} \geq (J^t(A^r \otimes B^r)J)^{1/r},$$

which is, by (2.3), inequality (2.1). □

Special cases. Some special cases of (2.1) are the following:

$$(A^{-1} \circ B^{-1})^{-1} \leq A \circ B \tag{2.5}$$

or, equivalently,

$$(A \circ B)^{-1} \leq A^{-1} \circ B^{-1}. \tag{2.6}$$

For positive integer r,

$$A \circ B \leq (A^r \circ B^r)^{1/r} \tag{2.7}$$

from which we can get

$$A^{1/r} \circ B^{1/r} \leq (A \circ B)^{1/r}. \tag{2.8}$$

These last two results are extensions of (1.2) and (1.3).

Remark. Inequalities (1.1) and (2.6) can be obtained by using Jensen's inequality for matrix convex functions, i.e., for matrix convex function f [6]

$$f(V^*AV) \leq V^*f(A)V. \tag{2.9}$$

Namely, using this result for the matrix convex function $f(t) = t^2$, we can get

$$\begin{aligned}(A \circ B)^2 = (J^t(A \otimes B)J)^2 &\leq J^t(A \otimes B)^2J \\ &= J^t(A^2 \otimes B^2)J = A^2 \circ B^2\end{aligned}$$

which is (1.1).

Similarly, we can use (2.9) for the matrix convex function $f(t) = t^{-1}$ to get (2.6).

THEOREM 2.2. *Let A and B be two positive definite $n \times n$ Hermitian matrices and let r and s be nonzero integers such that $r < s$. Then*

$$r(A^r \circ B^r - aA^s \circ B^s - bI) \geq 0, \tag{2.10}$$

where

$$a = (M^r - m^r)/(M^s - m^s), \quad b = (M^s m^r - M^r m^s)/(M^s - m^s),$$

and M and m are the largest and smallest eigenvalues of $A \otimes B$.

Proof. We have the matrix inequality [7]

$$r(A^r - aA^s - bI) \geq 0,$$

i.e.,

$$r[(A \otimes B)^r - a(A \otimes B)^s - bI] \geq 0.$$

Therefore, from (2.4),

$$r[A^r \otimes B^r - a(A^s \otimes B^s) - bI] \geq 0.$$

Now pre- and post-multiplication by J^t and J, respectively, give (2.10). □

Remark. We can also prove Theorem 2.2 by using Theorem 1 from [8].

THEOREM 2.3. *Let the conditions of Theorem 2.2 be satisfied. Then*

$$(A^s \circ B^s)^{1/s} \leq \tilde{\Delta}(A^r \circ B^r)^{1/r}, \tag{2.11}$$

where

$$\tilde{\Delta} = \left\{ \frac{r(\gamma^s - \gamma^r)}{(s-r)(\gamma^r - 1)} \right\}^{1/s} \quad \left\{ \frac{s(\gamma^r - \gamma^s)}{(r-s)(\gamma^s - 1)} \right\}^{-1/r} \tag{2.12}$$

and $\gamma = M/m$.

Proof. Let A be an $n \times n$ positive definite Hermitian matrix with eigenvalues contained in the interval $[m, M]$, where $0 < m < M$, and let V be an $n \times t$ matrix such that $V^*V = I$. If r, s are nonzero real numbers such that $s > r$ and either $s \notin (-1, 1)$ or $r \notin (-1, 1)$, then [8]

$$(V^* A^s V)^{1/s} \leq \tilde{\Delta}(V^* A^r V)^{1/r}, \tag{2.13}$$

where $\tilde{\Delta}$ is given by (2.12).

Therefore, in our case, we have

$$(A^s \circ B^s)^{1/s} = (J^t(A^s \otimes B^s)J)^{1/s} = (J^t(A \otimes B)^s J)^{1/s}$$
$$\leq \tilde{\Delta}(J^t(A \otimes B)^r J)^{1/r} = \tilde{\Delta}(J^t(A^r \otimes B^r)J)^{1/r} = \tilde{\Delta}(A^r \circ B^r)^{1/r}. \quad \square$$

Special cases.

1. For $s = 2$ and $r = 1$, we get (1.5).
2. For $s = 1,\ r = -1$, we get

$$A \circ B \leq \frac{(m+M)^2}{4Mm}(A^{-1} \circ B^{-1})^{-1} \tag{2.14}$$

or, equivalently,

$$A^{-1} \circ B^{-1} \leq \frac{(M+m)^2}{4Mm}(A \circ B)^{-1}. \tag{2.15}$$

THEOREM 2.4. *Let the conditions of Theorem* 2.2 *be satisfied. Then*

$$(A^s \circ B^s)^{1/s} - (A^r \circ B^r)^{1/r} \leq \Delta I, \tag{2.16}$$

where

$$\Delta = \max_{\theta \in [0,1]} \left\{ [\theta M^s + (1-\theta)m^s]^{1/s} - [\theta M^r + (1-\theta)m^r]^{1/r} \right\}. \tag{2.17}$$

Proof. Let A be an $n \times n$ positive definite Hermitian matrix with eigenvalues contained in the interval $[m, M]$, where $0 < m < M$, and let V be an $n \times t$ matrix such that $V^*V = I$. If $r,\ s$ are nonzero real numbers such that $s > r$ and either $s \notin (-1, 1)$ or $r \notin (-1, 1)$, then [8]

$$(V^*A^sV)^{1/s} - (V^*A^rV)^{1/r} \leq \Delta I, \tag{2.18}$$

where Δ is given by (2.17).

Thus, in our case, we have

$$(A^s \circ B^s)^{1/s} - (A^r \circ B^r)^{1/r} = [J^t(A^s \otimes B^s)]^{1/s}$$
$$- [J^t(A^r \otimes B^r)J]^{1/r} = [J^t(A \otimes B)^s J]^{1/s} - [J^t(A \otimes B)^r J]^{1/r} \leq \Delta I. \qquad \square$$

Special cases.

1. For $s = 2,\ r = 1$, we get

$$(A^2 \circ B^2)^{1/2} - A \circ B \leq \frac{(M-m)^2}{4(M+m)} I. \tag{2.19}$$

2. For $s = 1,\ r = -1$, we get

$$A \circ B - (A^{-1} \circ B^{-1})^{-1} \leq (\sqrt{M} - \sqrt{m})^2 I. \tag{2.20}$$

We note that the eigenvalues of $A \otimes B$ are the n^2 products of the eigenvalues of A by the eigenvalues of B [2, p. 245]. Thus if the eigenvalues of A and B, respectively, are ordered by

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n > 0, \qquad \beta_1 \geq \beta_2 \geq \cdots \geq \beta_n,$$

then in all previous results $M = \alpha_1\beta_1$ and $m = \alpha_n\beta_n$.

Thus (1.4), (1.5), (2.14), (2.13), (2.19), and (2.20) become, respectively,

$$A^2 \circ B^2 - (A \circ B)^2 \leq \frac{1}{4}(\alpha_1\beta_1 - \alpha_n\beta_n)^2 I,$$

$$(A^2 \circ B^2)^{1/2} \leq \frac{(\alpha_1\beta_1 + \alpha_n\beta_n)}{2\sqrt{\alpha_1\beta_1\alpha_n\beta_n}} A \circ B,$$

$$A \circ B \leq \frac{(\alpha_1\beta_1 + \alpha_n\beta_n)^2}{4\alpha_1\beta_1\alpha_n\beta_n}(A^{-1} \circ B^{-1})^{-1},$$

$$A^{-1} \circ B^{-1} \leq \frac{(\alpha_1\beta_1 + \alpha_n\beta_n)^2}{4\alpha_1\beta_1\alpha_n\beta_n}(A \circ B)^{-1},$$

$$(A^2 \circ B^2)^{1/2} - A \circ B \leq \frac{(\alpha_1\beta_1 - \alpha_n\beta_n)^2}{4(\alpha_1\beta_1 + \alpha_n\beta_n)} I,$$

and

$$A \circ B - (A^{-1} \circ B^{-1})^{-1} \leq (\sqrt{\alpha_1\beta_1} - \sqrt{\alpha_n\beta_n})^2 I.$$

REFERENCES

[1] R. A. Horn, *The Hadamard product*, in Matrix Theory and Applications, C. R. Johnson, ed., Proc. Sympos. Appl. Math., 40 (1989), pp. 87–169.

[2] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[3] T. Kollo and H. Neudecker, *Asymptotics of eigenvalues and unit-length eigenvectors of sample variance and correlation matrices*, J. Multivariate Anal., 47 (1993), pp. 283–300.

[4] S. Liu and H. Neudecker, *Several matrix Kantorovich-type inequalities*, J. Math. Anal. Appl., 197 (1996), pp. 23–26.

[5] B. Mond and J. E. Pečarić, *On Jensen's inequality for operator convex functions*, Houston J. Math., 21 (1995), pp. 739–754.

[6] D. Kainuma and M. Nakamura, *Around Jensen's inequality*, Math. Japon., 25 (1980), pp. 585–588.

[7] A. W. Marshall and I. Olkin, *Matrix versions of the Cauchy and Kantorovich inequalities*, Aequationes Math., 40 (1990), pp. 89–93.

[8] B. Mond and J. E. Pečarić, *A matrix version of the Ky Fan generalization of the Kantorovich inequality* II, Linear and Multilinear Algebra, 38 (1995), pp. 309–313.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 71–88, January 1998

005

PRIMITIVITY OF POSITIVE MATRIX PAIRS: ALGEBRAIC CHARACTERIZATION, GRAPH THEORETIC DESCRIPTION, AND 2D SYSTEMS INTERPRETATION*

ETTORE FORNASINI† AND MARIA ELENA VALCHER†

Abstract. In this paper the primitivity of a positive matrix pair (A, B) is introduced as a strict positivity constraint on the asymptotic behavior of the associated two-dimensional (2D) state model. The state evolution is first considered under the assumption of periodic initial conditions. In this case the system evolves according to a one-dimensional (1D) state updating equation, described by a block circulant matrix. Strict positivity of the asymptotic dynamics is equivalent to the primitivity of the circulant matrix, a property that can be restated as a set of conditions on the spectra of $A + e^{i\omega}B$, for suitable real values of ω.

The theory developed in this context provides a foundation whose analytical ideas may be generalized to nonperiodic initial conditions. To this purpose the spectral radius and the maximal modulus eigenvalues of the matrices $e^{i\theta}A + e^{i\omega}B$, θ and $\omega \in \mathbb{R}$, are related to the characteristic polynomial of the pair (A, B) as well as to the structure of the graphs associated with A and B and to the factorization properties of suitable integer matrices. A general description of primitive positive matrix pairs is finally derived, including both spectral and combinatorial conditions on the pair.

Key words. primitive matrices, circulant matrices, directed graphs, integer matrices, multidimensional systems

AMS subject classifications. 15A48, 11C20, 11A07, 15A18, 93C55

PII. S0895479895294095

1. Introduction.

The notion of primitive matrix grew out of the study of the spectra and the directed graphs of positive irreducible matrices, in a purely algebraic context [2], [3], [15]. Indeed, an irreducible matrix $F \in \mathbb{R}_+^{n\times n}$ is primitive if and only if its spectral radius is the only maximal modulus eigenvalue of F or, equivalently, if and only if in the associated directed graph the gcd of the lengths of all circuits is unitary.

An alternative definition of primitivity arises in the asymptotic analysis of the homogeneous discrete time positive system

$$\mathbf{x}(t+1) = F\mathbf{x}(t), \qquad t = 0, 1, \ldots, \tag{1}$$

when $\mathbf{x}(0)$, the initial state, is a nonnegative vector. Positive systems appear quite frequently in modeling real processes whose variables represent intrinsically nonnegative quantities, such as pressures, concentrations, densities, population levels, etc., and have been the object of a long stream of research aiming to explore basic issues of linear system theory, like controllability, reachability [4], [6], [7], [17], and realizability [1], [14] under positivity constraints. In this context, the primitivity of F can be equivalently restated as the property that every positive initial condition $\mathbf{x}(0)$ produces a state evolution which becomes strictly positive within a finite number of steps.

When trying to introduce a notion of primitivity for a positive matrix pair (A, B), with A and B in $\mathbb{R}_+^{n\times n}$, an extension of the above algebraic characterizations is not

* Received by the editors October 31, 1995; accepted for publication (in revised form) by P. Lancaster December 18, 1996.

http://www.siam.org/journals/simax/19-1/29409.html

† Dipartimento di Elettronica ed Informatica, Università di Padova, via Gradenigo 6a, 35131 Padova, Italy (fornasini@dei.unipd.it, meme@dei.unipd.it).

immediately apparent, whereas it is easy to figure out a reasonable extension of the dynamical behavior we have just described. To this end, we associate with the pair (A, B) the discrete homogeneous 2D system [8]

$$\mathbf{x}(h+1,k+1) = A\mathbf{x}(h,k+1) + B\mathbf{x}(h+1,k), \qquad h,k \in \mathbb{Z},\ h+k \geq 0, \tag{2}$$

where the doubly indexed *local states* $\mathbf{x}(h,k)$ are elements of the positive orthant $\mathbb{R}^n_+$ and *initial conditions* are given by assigning a sequence $\mathcal{X}_0 := \{\mathbf{x}(\ell,-\ell) : \ell \in \mathbb{Z}\}$ of nonnegative local states on the *separation set* $\mathcal{C}_0 := \{(\ell,-\ell) : \ell \in \mathbb{Z}\}$. A 2D system satisfying these constraints is called a 2D *positive system* [11].

2D state models described in (2) allow us to represent processes or devices whose evolutions depend upon two independent variables, according to a quarter plane causality law, and provide suitable descriptions of a large class of phenomena. They were introduced in the early seventies, and most of their internal and external features have been subsequently investigated. 2D positive systems, instead, have made their appearance only recently in some contributions dealing with the discretization of the set of PDEs describing a diffusion process [9], [12], but still their relevance for modeling certain classes of physical processes has been immediately apparent.

By assuming the aforementioned dynamical viewpoint, and in analogy with the 1D case, we express the primitivity of the pair (A, B) as a strict positivity constraint on the asymptotic behavior of (2). It is easy to see, however, that the structure of the sequence $\mathcal{X}_0$ has to be somehow constrained. In fact, if $\mathcal{X}_0$ includes $N+1$ consecutive zero local states

$$\mathbf{x}(h,-h) = \mathbf{x}(h+1,-h-1) = \cdots = \mathbf{x}(h+N,-h-N) = \mathbf{0},$$

then zero local states occur also on the separation sets

$$\mathcal{C}_t := \{(t+\ell,-\ell) : \ell \in \mathbb{Z}\}, \qquad t = 1,2,\ldots,N,$$

irrespective of the remaining initial conditions on $\mathcal{C}_0$. So, in order to guarantee that for some finite t all local states on $\mathcal{C}_t$ are strictly positive, we must restrict our attention to *admissible* sequences of initial conditions, namely, to nonnegative sequences $\mathcal{X}_0$ which satisfy the following assumption: there is an integer $N > 0$ such that $\sum_{\ell=h}^{h+N} \mathbf{x}(\ell,-\ell) > 0$ for all $h \in \mathbb{Z}$. We are now in a position to introduce the following definition of primitivity for a nonnegative matrix pair.

DEFINITION 1.1. *A pair of nonnegative matrices* (A, B) *is* primitive *if, for every admissible sequence* $\mathcal{X}_0$ *of initial conditions, all local states* $\mathbf{x}(h,k)$ *become strictly positive when* $h+k$ *is sufficiently large.*

Notice that when a 2D system is described by a primitive matrix pair, eventually all its variables appear "permanently excited," independent of the particular set of admissible initial conditions that originated its evolution. This seems to be particularly relevant when the system describes, for instance, a diffusion process and the two independent variables represent a spatial and a temporal coordinate. In that case, primitivity guarantees that, after a certain time instant, at every point all system variables represent strictly positive quantities.

To investigate the spectral and combinatorial properties of a primitive matrix pair, we consider first the dynamics of system (2) when the initial conditions sequence $\mathcal{X}_0$ has a periodic pattern of period T. Under this assumption, the 2D system exhibits a behavior which is somewhat intermediate between those of (1) and (2), as its state

evolution can be equivalently described by a model (1) with an $nT \times nT$ block circulant system matrix $F = C_T(A, B)$. It is clear that $\mathbf{x}(h, k)$ eventually becomes strictly positive if and only if $C_T(A, B)$ is primitive, a property that easily translates into the condition that the spectral radii of the matrices $A + e^{i2\pi\ell/T}B,\ \ell = 1, \ldots, T-1$, are smaller than the spectral radius of $A + B$.

So, the primitivity of all circulant matrices $C_T(A, B),\ T = 1, 2, \ldots$, which is a necessary condition for the primitivity of (A, B), is equivalent to assuming that $A + e^{i\omega}B$ has spectral radius smaller than that of $A + B$, whenever ω is a rational, but not an integer, multiple of 2π.

This remark suggests a way for obtaining equivalent descriptions of the primitivity of (A, B), based on the spectral properties of the matrix family $\{A + e^{i\omega}B : \omega \in \mathbb{R}\}$. Actually, searching for a graph theoretic interpretation of the primitivity condition of all circulant matrices $C_T(A, B)$, we can show that it corresponds to simple constraints on the structure of a certain directed graph $\mathcal{D}^*(A, B)$ associated with the pair (A, B), and on the integer matrix $L_{A,B}$ which describes its cyclic structure. Tying together these combinatorial characterizations with a result [10] on the Hurwitz products involved in the state updating of (2), we prove that the primitivity of all $C_T(A, B)$, $T = 1, 2, \ldots$, is also sufficient for that of the pair (A, B).

The paper is organized as follows: the next section investigates the spectral and combinatorial features of the pair (A, B) by means of the complex matrices $e^{i\theta}A + e^{i\omega}B$, $\theta, \omega \in \mathbb{R}$, and of the directed graph $\mathcal{D}^*(A, B)$, respectively. Section 3 analyzes the periodic dynamics of system (2) and the properties of the associated circulant matrices $C_T(A, B),\ T = 1, 2, \ldots$. Finally, in section 4, the primitivity of (A, B) is shown to be equivalent to a set of conditions involving the cyclic structure of $\mathcal{D}^*(A, B)$, the spectra of $e^{i\theta}A + e^{i\omega}B$, $\theta, \omega \in \mathbb{R}$, and the positivity of at least one Hurwitz product.

As we assume familiarity with the basic results of graph theory and positive matrix theory, they will be only touched upon in this introduction to explain the notation in use throughout the paper. Although some elementary background on 2D systems will be provided later in this section, a couple of algebraic facts will be stated without proof. The interested reader is referred to [10], which includes further references on the subject.

Matrices and vectors will usually be represented by capital italic and lower case boldface letters, respectively, while their entries are represented by the corresponding lower case italic letters. Sometimes, however, when a matrix F is expressed as the product or the sum of other matrices, it will be convenient to denote its (i, j)th entry as $[F]_{ij}$. If $F = [f_{ij}]$ is a matrix (in particular, a vector), we write $F \gg 0$ (F *strictly positive*), if $f_{ij} > 0$ for all i, j; $F > 0$ (F *positive*), if $f_{ij} \geq 0$ for all i, j, and $f_{hk} > 0$ for some pair (h, k); $F \geq 0$ (F *nonnegative*), if $f_{ij} \geq 0$ for all i, j. The *spectral radius* of a matrix F, i.e., the modulus of its maximal eigenvalue, is denoted by $\rho(F)$. Every $n \times n$ nonnegative matrix F has a corresponding [3] *digraph* (directed graph) $\mathcal{D}(F)$ of *order* n, with vertices indexed by $1, 2, \ldots, n$. There is an arc (i, j) from i to j if and only if $f_{ij} > 0$. Similarly, we associate with a pair of $n \times n$ nonnegative matrices (A, B) a digraph of order n, $\mathcal{D}^*(A, B)$, with arcs of two different kinds, namely, A-arcs and B-arcs. There is an A-arc from vertex i to vertex j if and only if $a_{ij} > 0$, and a B-arc if and only if $b_{ij} > 0$.

Example 1. Consider the pair of positive matrices

$$A = \begin{bmatrix} 0 & 0 & 2 \\ 1 & 4 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & 6 \\ 0 & 0 & 0 \\ 0 & 4 & 0 \end{bmatrix}.$$

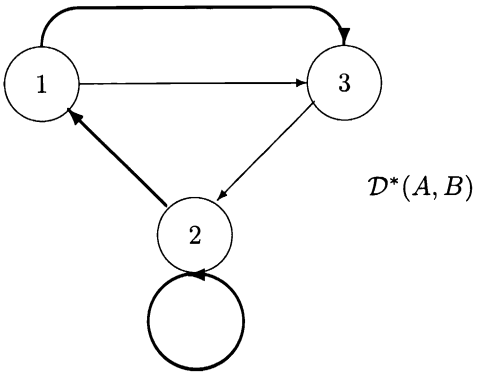


FIG. 1.

The associated digraph $\mathcal{D}^*(A,B)$ is given in Fig. 1, where A-arcs and B-arcs have been represented (as in what follows) by thick lines and thin lines, respectively.

A sequence of arcs in $\mathcal{D}(F)$ of the form $(i_0,i_1),(i_1,i_2),\ldots,(i_{k-1},i_k)$ defines a *path of length* k in $\mathcal{D}(F)$, connecting i_0 to i_k. When assigning a path p in $\mathcal{D}^*(A,B)$, we also have to specify, for each pair of consecutive vertices, which kind of arc they are connected by, so that p will have a representation like $(i_0,i_1)_A,(i_1,i_2)_B,\ldots,(i_{k-1},i_k)_B$. Thus, it is natural to associate p with a couple of nonnegative integers, $\alpha(p)$ and $\beta(p)$, representing the number of A-arcs and B-arcs occurring in p, respectively. A path whose extreme vertices coincide, i.e., $i_0=i_k$, is called a *cycle*. In particular, if each vertex in a cycle appears exactly once as the first vertex of an arc, the cycle is called a *circuit*.

Given a pair of square matrices (A,B), not necessarily nonnegative, the *Hurwitz products* of A and B are inductively defined [10] as

$$A^i \sqcup\!\sqcup^0 B = A^i, \quad i \geq 0, \qquad \text{and} \qquad A^0 \sqcup\!\sqcup^j B = B^j, \quad j \geq 0, \tag{3}$$

and, when i and j are both greater than zero,

$$A^i \sqcup\!\sqcup^j B = A(A^{i-1} \sqcup\!\sqcup^j B) + B(A^i \sqcup\!\sqcup^{j-1} B). \tag{4}$$

One easily sees that $A^i \sqcup\!\sqcup^j B$ is the sum of all matrix products that include the factors A and B, i and j times, respectively. For notational convenience sometimes we allow either i or j to be negative integers, and in these cases we assume $A^i \sqcup\!\sqcup^j B = 0$. Hurwitz products allow us to express any local state $\mathbf{x}(h,k)$ of system (2) in terms of the sequence of initial conditions. Actually, if $\mathcal{X}_0 = \{\mathbf{x}(\ell,-\ell) : \ell \in \mathbb{Z}\}$ is an arbitrary sequence of initial conditions on $\mathcal{C}_0$, for all $h,k \in \mathbb{Z}$, $h+k \geq 0$, $\mathbf{x}(h,k)$ can be represented as

$$\mathbf{x}(h,k) = \sum_{\ell} (A^{h-\ell} \sqcup\!\sqcup^{k+\ell} B)\ \mathbf{x}(\ell,-\ell). \tag{5}$$

In particular, if the initial conditions on the separation set $\mathcal{C}_0$ are all zero, except at $(0,0)$, we have

$$\mathbf{x}(h,k) = (A^h \sqcup\!\sqcup^k B)\ \mathbf{x}(0,0) \qquad \forall\ h,k \geq 0.$$

The *characteristic polynomial* of a pair of $n \times n$ matrices (A, B) is defined as

$$\Delta_{A,B}(z_1, z_2) := \det(I_n - Az_1 - Bz_2)$$

and plays for system (2) the same role as $\det(I_n - Fz)$ for system (1). In particular, there is a bijective correspondence [10] between the characteristic polynomial of a pair (A, B) and the family of traces $\mathrm{tr}(A^{i} \sqcup\!\sqcup^{j} B)$, $i, j \in \mathbb{N}$, a result which generalizes the well-known relation [13] between the coefficients of $\det(I_n - Fz)$ and the traces of all powers of F.

2. Spectral properties of the matrices $e^{i\theta}A + e^{i\omega}B$. The Perron–Frobenius theory establishes, for an $n \times n$ irreducible matrix F, very tight connections among its characteristic polynomial, the invariance under rotation of its spectrum, and the lengths of all cycles in the associated digraph $\mathcal{D}(F)$. These connections can be specialized to primitive matrices, thus leading to a set of characterizations of primitivity which represent suitable strengthenings of those available for irreducibility. Trying to determine necessary and sufficient conditions for the primitivity of a positive matrix pair (A, B), it seems natural to ask to what extent the above results admit a generalization, once the spectrum of F is replaced by the variety of $\Delta_{A,B}(z_1, z_2)$ and the digraph of F by $\mathcal{D}^*(A, B)$. To this purpose, in this section and throughout the paper, we will steadily assume that the matrix pair (A, B) we are considering has the following properties:

a) A and B are both positive;
b) $A + B$ is irreducible;
c) $A + B$ has a unitary maximal eigenvalue.

The set of $n \times n$ pairs endowed with these properties will be denoted by $\mathcal{I}_n$. Assumptions a) and b) easily prove to be necessary conditions for 2D primitivity, which is our final goal. Actually, requiring that all states on $\mathcal{C}_t$ are strictly positive for large values of t implies that both A and B are nonzero, otherwise any sequence $\mathcal{X}_0$ including a zero local state would produce on every $\mathcal{C}_t$ a state sequence with the same property. Analogously, if $A + B$ were reducible, a positive, but not strictly positive, vector $\mathbf{c}$ could be found such that the initial state sequence $\mathcal{X}_0 = \{\mathbf{x}(\ell, -\ell) = \mathbf{c} : \forall\ \ell \in \mathbb{Z}\}$ produces a constant sequence of nonstrictly positive local states on every separation set $\mathcal{C}_t$. Assumption c) entails no loss of generality. Actually, we can divide both A and B by $\rho(A + B)$ without affecting the properties we aim to investigate, which are independent of the spectral radius of $A + B$. The case when $A + B$ is nilpotent would constitute the unique exception to this rescaling procedure, but then $A + B$ would not be irreducible.

The answer to the previous question is given by the following proposition, which enlightens, under different points of view, which rotations of θ and ω radians in the z_1- and z_2-planes, respectively, leave the variety of $\Delta_{A,B}(z_1, z_2)$ invariant. The proof is based on the following remarkable result due to Wielandt [15].

Wielandt's theorem. If an $n \times n$ complex matrix $C = [c_{ij}]$ is dominated by an irreducible matrix $F = [f_{ij}] > 0$, i.e., $|c_{ij}| \le f_{ij}$, for all i and j, then for all eigenvalues λ_C of C

$$|\lambda_C| \le \rho(F). \tag{6}$$

Equality holds in (6) if and only if

$$C = e^{i\phi} D F D^{-1}, \tag{7}$$

where $\lambda_C = e^{i\phi}\rho(F)$ and $D = \text{diag}\{e^{i\omega_1}, e^{i\omega_2}, \dots, e^{i\omega_n}, \}$, $\omega_1, \omega_2, \dots, \omega_n \in \mathbb{R}$.

PROPOSITION 2.1. *Let $(A, B) \in \mathcal{I}_n$. For any θ and $\omega \in \mathbb{R}$ the following facts are equivalent:*

i) 1 *is an eigenvalue of* $e^{i\theta}A + e^{i\omega}B$;

ii) *there exists a diagonal matrix* $D = \text{diag}\{e^{i\omega_1}, e^{i\omega_2}, \dots, e^{i\omega_n}, \}$, $\omega_1, \omega_2, \dots, \omega_n \in \mathbb{R}$, *such that*

$$A = e^{i\theta}DAD^{-1} \qquad \text{and} \qquad B = e^{i\omega}DBD^{-1}; \tag{8}$$

iii) *for every cycle γ in $\mathcal{D}^*(A, B)$, including $\alpha(\gamma)$ A-arcs and $\beta(\gamma)$ B-arcs,*

$$\alpha(\gamma)\theta + \beta(\gamma)\omega \equiv 0 \qquad \text{mod } 2\pi; \tag{9}$$

iv) *the characteristic polynomial of the pair (A, B) satisfies*

$$\Delta_{A,B}(z_1, z_2) = \Delta_{A,B}(z_1 e^{i\theta}, z_2 e^{i\omega}). \tag{10}$$

Proof. i) $\Rightarrow$ ii) As the matrix $e^{i\theta}A + e^{i\omega}B$ is dominated by $A + B$ and condition i) holds, by Wielandt's theorem we have $\rho(e^{i\theta}A + e^{i\omega}B) = \rho(A + B) = 1$ and

$$A + B = D(e^{i\theta}A + e^{i\omega}B)D^{-1}, \tag{11}$$

for some diagonal matrix $D = \text{diag}\{e^{i\omega_1}, e^{i\omega_2}, \dots, e^{i\omega_n}, \}$, $\omega_1, \omega_2, \dots, \omega_n \in \mathbb{R}$. If $a_{hk} \neq 0$, from (11) one gets

$$e^{i\omega_h}(e^{i\theta}a_{hk} + e^{i\omega}b_{hk})e^{-i\omega_k} = a_{hk} + b_{hk},$$

and consequently

$$(1 - e^{i(\theta+\omega_h-\omega_k)})a_{hk} = -(1 - e^{i(\omega+\omega_h-\omega_k)})b_{hk}. \tag{12}$$

As the real parts on the left and right sides of (12) are nonnegative and nonpositive, respectively, they must be zero, and hence $\omega_k \equiv \omega_h + \theta \mod 2\pi$. So, we have

$$[e^{i\theta}DAD^{-1}]_{hk} = e^{i\theta}e^{i\omega_h}a_{hk}e^{-i\omega_k} = a_{hk},$$

which proves the first equation in (8). The second one immediately follows from (11).

ii) $\Rightarrow$ iii) Let $\gamma = (g_1, g_2), \dots, (g_{\ell-1}, g_\ell), (g_\ell, g_1)$ be a cycle of length ℓ in $\mathcal{D}^*(A, B)$, including $\alpha(\gamma)$ A-arcs and $\beta(\gamma)$ B-arcs. For every arc (g_i, g_j) in γ, let $c_{g_i g_j}$ denote $a_{g_i g_j}$ if (g_i, g_j) is an A-arc and $b_{g_i g_j}$ if it is a B-arc. By (8) we have, then,

$$0 < c_{g_1 g_2}c_{g_2 g_3} \dots c_{g_\ell g_1} = e^{i[\alpha(\gamma)\theta+\beta(\gamma)\omega]}c_{g_1 g_2}c_{g_2 g_3} \dots c_{g_\ell g_1},$$

which implies (9).

iii) $\Rightarrow$ iv) Consider any Hurwitz product $A^h \sqcup\!\sqcup^k B$, with $h, k \in \mathbb{N}$, $h + k > 0$. If $\text{tr}(A^h \sqcup\!\sqcup^k B) \neq 0$, there is a circuit γ in $\mathcal{D}^*(A, B)$, including h A-arcs and k B-arcs and, by assumption, the congruence relation $h\theta + k\omega \equiv 0 \bmod 2\pi$ is satisfied. Consequently, the identity $\text{tr}(A^h \sqcup\!\sqcup^k B)[1 - e^{i(h\theta+k\omega)}] = 0$ holds for all integers h and k, and we get

$$\text{tr}(A^h \sqcup\!\sqcup^k B) = e^{i(h\theta+k\omega)}\text{tr}(A^h \sqcup\!\sqcup^k B) = \text{tr}\Big((e^{i\theta}A)^h \sqcup\!\sqcup^k (e^{i\omega}B)\Big). \tag{13}$$

As the traces of the Hurwitz products uniquely determine the coefficients of the characteristic polynomial of a matrix pair [10], it follows that

$$\Delta_{A,B}(z_1, z_2) = \det(I - Az_1 - Bz_2) = \det\Big(I - (e^{i\theta}A)z_1 - (e^{i\omega}B)z_2\Big) = \Delta_{A,B}(z_1 e^{i\theta}, z_2 e^{i\omega}).$$

iv) $\Rightarrow$ i) As the pair (A, B) is in $\mathcal{I}_n$, 1 is an eigenvalue of $A + B$. Consequently,

$$0 = \det(I - A - B) = \Delta_{A,B}(1, 1) = \Delta_{A,B}(e^{i\theta}, e^{i\omega}) = \det(I - e^{i\theta}A - e^{i\omega}B),$$

which implies $1 \in \Lambda(e^{i\theta}A + e^{i\omega}B)$. □

Remarks. a) In order to check condition iii) of Proposition 2.1, it is not necessary to consider all cycles but only the circuits in $\mathcal{D}^*(A, B)$. So, point iii) reduces to a finite number, say t, of congruence relations which can be expressed in matrix form as

$$L_{A,B}\begin{bmatrix}\theta\\ \omega\end{bmatrix} = \begin{bmatrix}\alpha(\gamma_1) & \beta(\gamma_1)\\ \alpha(\gamma_2) & \beta(\gamma_2)\\ \vdots & \vdots\\ \alpha(\gamma_t) & \beta(\gamma_t)\end{bmatrix}\begin{bmatrix}\theta\\ \omega\end{bmatrix} \equiv \mathbf{0} \qquad \text{mod } 2\pi. \tag{14}$$

b) If in $\mathcal{D}^*(A, B)$ both an A-arc and a B-arc can be found, connecting a vertex h to a vertex k, there are two cycles γ_1 and γ_2 with $\alpha(\gamma_2) = \alpha(\gamma_1)-1$ and $\beta(\gamma_2) = \beta(\gamma_1)+1$. As the pairs (θ, ω) which satisfy (9) for all γ in $\mathcal{D}^*(A, B)$ must, in particular, satisfy

$$\begin{aligned}\alpha(\gamma_1)\theta + \beta(\gamma_1)\omega &\equiv 0 \qquad \text{mod } 2\pi,\\ \Big(\alpha(\gamma_1) - 1\Big) + \Big(\beta(\gamma_1) + 1\Big)\omega &\equiv 0 \qquad \text{mod } 2\pi,\end{aligned}$$

we have $\theta \equiv \omega \bmod 2\pi$ for all solutions of (9).

c) Finally, notice that condition $1 \in \Lambda(e^{i\theta}A + e^{i\omega}B)$ for some real pair (θ, ω) is equivalent to the fact that, for a suitable real pair (ϕ, ψ), $e^{i\phi}$ is an eigenvalue of $A + e^{i\psi}B$.

If (A, B) is an element of $\mathcal{I}_n$ and $A + B$ is primitive, the situation when only the trivial rotations, i.e., $\theta \equiv \omega \equiv 0 \bmod 2\pi$, leave invariant the variety of $\Delta_{A,B}(z_1, z_2)$, corresponds to the special case when the congruence (14) is devoid of nonzero solutions. This happens if and only if $L_{A,B}$ is a right prime integer matrix.

PROPOSITION 2.2. *Let $(A, B) \in \mathcal{I}_n$ and assume that $A + B$ is primitive. The following facts are equivalent:*

i) $1 \in \Lambda(e^{i\theta}A + e^{i\omega}B)$ *implies* $\theta \equiv \omega \equiv 0 \bmod 2\pi$;

ii) *the integer matrix $L_{A,B}$ is right prime*;

iii) $\Delta_{A,B}(z_1, z_2) = \Delta_{A,B}(e^{i\theta}z_1, e^{i\omega}z_2)$ *implies* $\theta \equiv \omega \equiv 0 \bmod 2\pi$.

Proof. i) $\Rightarrow$ ii) We show first that $L_{A,B}$ has full column rank. Consider the integer matrix

$$\bar{L}_{A,B} := \begin{bmatrix}\alpha(\gamma_1) & \alpha(\gamma_1) + \beta(\gamma_1)\\ \alpha(\gamma_2) & \alpha(\gamma_2) + \beta(\gamma_2)\\ \vdots & \vdots\\ \alpha(\gamma_t) & \alpha(\gamma_t) + \beta(\gamma_t)\end{bmatrix} = L_{A,B}\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix},$$

whose second column consists of the lengths of all circuits in $\mathcal{D}^*(A, B)$. The primitivity assumption on $A + B$ implies that the gcd of these lengths is 1, and hence integer coefficients x_h can be found such that $\sum_h x_h\,[\alpha(\gamma_h) + \beta(\gamma_h)] = 1$. If $\bar{L}_{A,B}$ were not full column rank, its first column, which is nonzero as (A, B) is in $\mathcal{I}_n$, would be a scalar multiple of the second one, namely,

$$\bar{L}_{A,B}\begin{bmatrix}1\\ -q\end{bmatrix} = \mathbf{0},$$

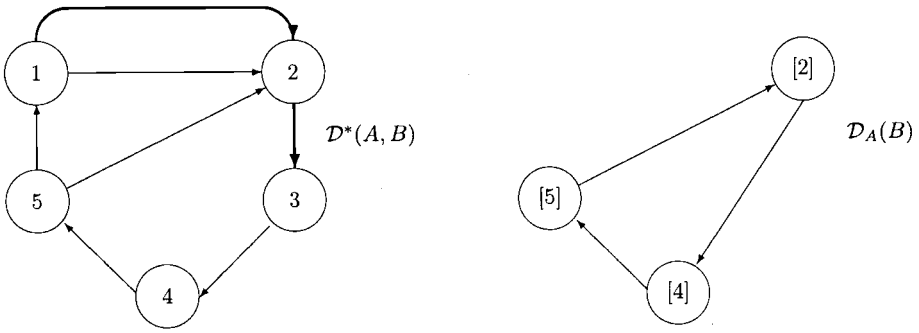


FIG. 2.

for some rational number q, $0 < q < 1$. Consequently, we would have $0 < \sum_h x_h \alpha(\gamma_h) < 1$, which is impossible, as all addenda $x_h\alpha(\gamma_h)$ are integer numbers. So, $L_{A,B}$, and hence $L_{A,B}$, have rank 2.

We prove now that $L_{A,B}$ is right prime. If not, it would factor over the ring $\mathbb{Z}$ as $L_{A,B} = L\Delta$, where L is a $t \times 2$ right prime matrix and Δ a square matrix with $\det \Delta \neq \pm 1$ [16]. As Δ^{-1} is not an integer matrix, the pair

$$\begin{bmatrix} \theta \\ \omega \end{bmatrix} := \Delta^{-1} \begin{bmatrix} 2\pi \\ 2\pi \end{bmatrix} \not\equiv \mathbf{0} \mod 2\pi$$

satisfies $L_{A,B}\left[\begin{smallmatrix} \theta \\ \omega \end{smallmatrix}\right] \equiv \mathbf{0} \mod 2\pi$. By Proposition 2.1, this implies $1 \in \Lambda(e^{i\theta}A + e^{i\omega}B)$.

ii) $\Rightarrow$ i) If $L_{A,B}$ is right prime, it admits a $2 \times t$ integer left inverse S, so that $SL_{A,B} = I_2$. Consequently, $L_{A,B}\left[\begin{smallmatrix} \theta \\ \omega \end{smallmatrix}\right] \equiv \mathbf{0} \mod 2\pi$ implies $\left[\begin{smallmatrix} \theta \\ \omega \end{smallmatrix}\right] \equiv \mathbf{0} \mod 2\pi$. By Proposition 2.1, this proves the result.

i) $\Leftrightarrow$ iii) This is obvious from Proposition 2.1. □

The situation when in θ is zero in Proposition 2.1 is particularly interesting for the subsequent analysis of circulant matrices. Clearly, the problem of determining for which ω's the matrix $A + e^{i\omega}B$ has eigenvalue 1 can be solved by resorting to the above propositions and, in particular, by analyzing the cyclic structure of $\mathcal{D}^*(A,B)$. It seems more convenient, however, to associate with the pair (A,B) a simpler (strongly connected) digraph $\mathcal{D}_A(B)$ obtained as follows: for all vertices $h \in \{1,2,\ldots,n\}$ shrink into a single vertex $[h]$ all vertices of the communicating class of h in $\mathcal{D}(A + A^T)$, and then connect $[h]$ and $[k]$ with the arc $([h],[k])$ if there is an arc (ℓ, m) in $\mathcal{D}(B)$ for some $\ell \in [h]$ and $m \in [k]$. The structure of the shrunken digraph $\mathcal{D}_A(B)$ of a pair (A,B) is better clarified by means of an example.

Example 2. The positive matrices

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 3 \\ 4 & 0 & 0 & 0 & 0 \end{bmatrix}$$

are associated with Fig. 2.

PROPOSITION 2.3. *Let $(A,B) \in \mathcal{I}_n$. 1 is an eigenvalue of $A + e^{i\omega}B$ if and only if the imprimitivity index $h_A(B)$ of the digraph $\mathcal{D}_A(B)$ satisfies*

$$h_A(B)\ \omega \equiv 0 \qquad \mod 2\pi. \tag{15}$$

Proof. By Proposition 2.1, the statement $1 \in \Lambda(A+e^{i\omega}B)$ can be replaced by the equivalent condition $\beta(\gamma)\ \omega \equiv 0 \mod 2\pi$, where γ ranges over all cycles in $\mathcal{D}^*(A,B)$ and $\beta(\gamma)$ denotes the number of B-arcs in γ. Assume, first, that $h_A(B)\ \omega \equiv 0 \mod 2\pi$. Every cycle γ in $\mathcal{D}^*(A,B)$, including, say, $\beta(\gamma)$ B-arcs, obviously determines a cycle γ' of length $\beta(\gamma)$ in $\mathcal{D}_A(B)$. As $h_A(B)$ is the gcd of the lengths of all cycles in $\mathcal{D}_A(B)$, the length $\beta(\gamma)$ of γ' satisfies $\beta(\gamma)\ \omega \equiv 0 \mod 2\pi$. To prove the converse, consider any cycle $\hat{\gamma}$ in $\mathcal{D}_A(B)$ of length, say, ℓ. By definition of $\mathcal{D}_A(B)$, there is a cycle $\bar{\gamma}$ in $\mathcal{D}^*(A+A^T,B)$ such that $\hat{\gamma}$ is obtained by identifying every pair of consecutive vertices connected in $\bar{\gamma}$ by an $(A+A^T)$-arc. As $A+B$ is irreducible, every A^T-arc (h,k) in $\bar{\gamma}$ can be replaced in $\mathcal{D}^*(A,B)$ by a suitable path p_{hk}, from h to k, thus producing a new cycle γ^*. Clearly, as $a_{kh}>0$, p_{hk} can be completed into a cycle γ_{hk} of $\mathcal{D}^*(A,B)$ by means of the A-arc corresponding to a_{kh}. Since all cycles γ_{hk} as well as γ^* satisfy

$$\begin{aligned} \beta(\gamma_{hk})\ \omega &\equiv 0 \qquad \mod 2\pi, \\ \beta(\gamma^*)\ \omega \equiv \Big(\ell + \sum \beta(\gamma_{hk})\Big)\ \omega &\equiv 0 \qquad \mod 2\pi, \end{aligned}$$

it follows that the length ℓ of any cycle in $\mathcal{D}_A(B)$ satisfies $\ell\omega \equiv 0 \mod 2\pi$, and hence $h_A(B)\omega \equiv 0 \mod 2\pi$. □

The results obtained in Proposition 2.2 for the linear combinations $e^{i\theta}A + e^{i\omega}B$ of the matrices A and B particularize to the case $\theta = 0$, by resorting once again to the shrunken digraph $\mathcal{D}_A(B)$.

PROPOSITION 2.4. *Let $(A,B) \in \mathcal{I}_n$. The following facts are equivalent:*

i) $1 \in \Lambda(A+e^{i\omega}B)$ *for some real number* ω *implies* $\omega \equiv 0 \mod 2\pi$;

ii) *gcd* $\{\beta(\gamma) : \gamma$ *a cycle in* $\mathcal{D}^*(A,B)\} = 1$;

iii) *the imprimitivity index* $h_A(B)$ *of* $\mathcal{D}_A(B)$ *is* 1.

Proof. i) ⇒ ii) If $b := \gcd\{\beta(\gamma) : \gamma$ a cycle in $\mathcal{D}^*(A,B)\}$ is greater than 1, then $\bar{\omega} := 2\pi/b$ is not an integer multiple of 2π. However, condition $\beta(\gamma)\ \ \bar{\omega} \equiv 0 \mod 2\pi$ holds true for every cycle γ in $\mathcal{D}^*(A,B)$, thus implying, by Proposition 2.1, that 1 is an eigenvalue of $A+e^{i\bar{\omega}}B$. This contradicts i).

ii) ⇒ iii) Given any cycle γ in $\mathcal{D}^*(A,B)$ with, say, $\beta(\gamma)$ B-arcs, we can identify pairs of consecutive vertices which are connected by A-arcs, thus obtaining a cycle in $\mathcal{D}_A(B)$ of length $\beta(\gamma)$. So, as gcd $\{\beta(\gamma) : \gamma$ a cycle in $\mathcal{D}^*(A,B)\} = 1$, there is a family of cycles in $\mathcal{D}_A(B)$ whose lengths are coprime, and hence $h_A(B)$ is 1.

iii) ⇒ i) This follows from Proposition 2.3. □

Remark. Analogous results can be obtained for the family of matrices $e^{i\theta}A + B$, $\theta \in \mathbb{R}$, by simply referring to the shrunken digraph $\mathcal{D}_B(A)$ and to the occurrences of the A-arcs in the cycles of $\mathcal{D}^*(A,B)$. It is worthwhile to notice, however, that the digraphs $\mathcal{D}_A(B)$ and $\mathcal{D}_B(A)$ can be endowed with different structural properties and, in particular, their imprimitivity indices $h_A(B)$ and $h_B(A)$ need not coincide.

To conclude this section we investigate the set of solutions of the congruence relation (14). As the pair (A,B) is in $\mathcal{I}_n$, both columns of $L_{A,B}$ are nonzero, and therefore we can distinguish two cases, depending on the rank of $L_{A,B}$.

• $L_{A,B}$ *has rank* 1 if and only if there is a pair of positive coprime integers, m and ℓ, such that

$$L_{A,B}\begin{bmatrix} m \\ -\ell \end{bmatrix} = \mathbf{0}.$$

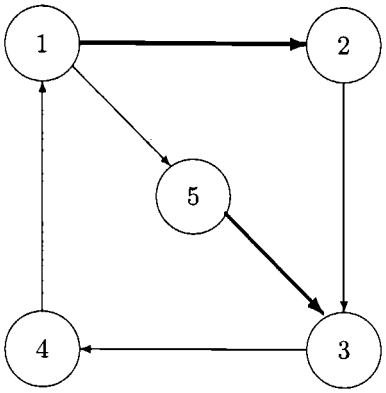


FIG. 3.

By the same reasoning adopted to prove Proposition 2.1, we see that the traces of the Hurwitz products $A^h \sqcup\!\sqcup^k B$ are possibly nonzero only for $(h,k) = (t\ell, tm)$, $t \in \mathbb{N}$. This situation corresponds [10] to a characteristic polynomial of the form $\Delta_{A,B}(z_1, z_2) = p(z_1^\ell z_2^m)$, i.e., with support included in a straight line through the origin. In this case the set of all distinct solutions of (14), i.e., corresponding to different pairs $(e^{i\theta}, e^{i\omega})$, includes infinitely many elements.

Example 3. The pair of matrices (A, B) in $\mathcal{I}_5$, with

$$A = \frac{1}{\sqrt[4]{2}} \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}, \qquad B = \frac{1}{\sqrt[4]{2}} \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

has characteristic polynomial $\Delta_{A,B}(z_1, z_2) = 1 - z_1 z_2^3 \in \mathbb{R}[z_1 z_2^3]$ and the associated digraph $\mathcal{D}^*(A, B)$ is shown in Fig. 3.

Clearly,

$$L_{A,B} = \begin{bmatrix} 1 & 3 \\ 1 & 3 \end{bmatrix}$$

has rank 1.

• When the support of $\Delta_{A,B}(z_1, z_2)$ is not included in a straight line, $L_{A,B}$ *has rank* 2 and there is only a finite set of distinct solutions of (14). To study this set, it is convenient to consider the Smith form of $L_{A,B}$ over $\mathbb{Z}$, namely,

$$S_{AB} = \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} = U L_{A,B} V,$$

where U and V are unimodular integer matrices, and the positive integers s_1 and $s_1 s_2$ represent the gcd's of the elements and of the second-order minors of $L_{A,B}$,

respectively. Equation (14) can be rewritten as

$$\begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} V^{-1} \begin{bmatrix} \theta \\ \omega \end{bmatrix} \equiv \mathbf{0} \mod 2\pi, \tag{16}$$

and hence as

$$\begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \equiv \mathbf{0} \mod \mathbb{Z}, \tag{17}$$

where $[\, r_1 \quad r_2 \,]^T := V^{-1} [\, \theta/2\pi \quad \omega/2\pi \,]^T$. A set of distinct representatives of all solutions of (17) is given by

$$\left\{ \begin{bmatrix} \frac{n_1}{s_1} \\ \frac{n_2}{s_2} \end{bmatrix} : n_1 = 0, 1, \ldots, s_1 - 1;\ n_2 = 0, 1, \ldots, s_2 - 1 \right\}.$$

So, letting $\mathbf{g}_1 := 2\pi V [\, 1/s_1 \quad 0 \,]^T$ and $\mathbf{g}_2 := 2\pi V [\, 0 \quad 1/s_2 \,]^T$, the set

$$\{n_1\mathbf{g}_1 + n_2\mathbf{g}_2 : n_1 = 0, 1, \ldots, s_1 - 1;\ n_2 = 0, 1, \ldots, s_2 - 1\}, \tag{18}$$

is the abelian group of the solutions (mod 2π) of (14), represented as the direct sum of two cyclic groups. The case when both cyclic groups are nontrivial is quite special because it occurs only when all elements of $L_{A,B}$ have a nontrivial common divisor s_1. In terms of Hurwitz products, this amounts to requiring that tr $(A^h \sqcup\!\sqcup^k B)$ is possibly nonzero only when both h and k are multiples of s_1 or, equivalently [10], $\Delta_{A,B}(z_1, z_2)$ is in $\mathbb{R}[z_1^{s_1}, z_2^{s_1}]$.

Example 4. The pair of positive matrices $(A, B) \in \mathcal{I}_5$, with

$$A = \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{3} & 0 & 0 \\ 0 & 0 & 0 & \sqrt{3} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

has characteristic polynomial $\Delta_{A,B}(z_1, z_2) = 1 - \frac{1}{4}z_1^2 - \frac{3}{4}z_1^2 z_2^2 \in \mathbb{R}[z_1^2, z_2^2]$. The associated digraph $\mathcal{D}^*(A, B)$ is shown in Fig. 4.

All entries of the matrix

$$L_{A,B} = \begin{bmatrix} 2 & 0 \\ 2 & 2 \end{bmatrix}$$

are multiples of 2.

In the remaining cases and, in particular, when $(A, B) \in \mathcal{I}_n$ has primitive sum $A + B$, the set of distinct solutions of (14) is a cyclic group generated by $\mathbf{g}_2$ and including s_2 elements. Finally, when $L_{A,B}$ *is right prime*, both cyclic groups collapse and we have only the trivial solution $\theta \equiv \omega \equiv 0 \mod 2\pi$.

As a final remark, if our interest is in the pairs $(0, \omega)$ which satisfy $L_{A,B}[\begin{smallmatrix} 0 \\ \omega \end{smallmatrix}] \equiv \mathbf{0}$ mod 2π or, equivalently, in the values $\omega \in [0, 2\pi[$ for which

$$1 \in \Lambda(A + e^{i\omega}B), \tag{19}$$

it is more convenient to exploit condition $h_A(B)\ \omega \equiv 0 \mod 2\pi$, given in Proposition 2.3. This way it is immediately apparent that the solutions (mod 2π) constitute a cyclic group of order $h_A(B) \leq n$.

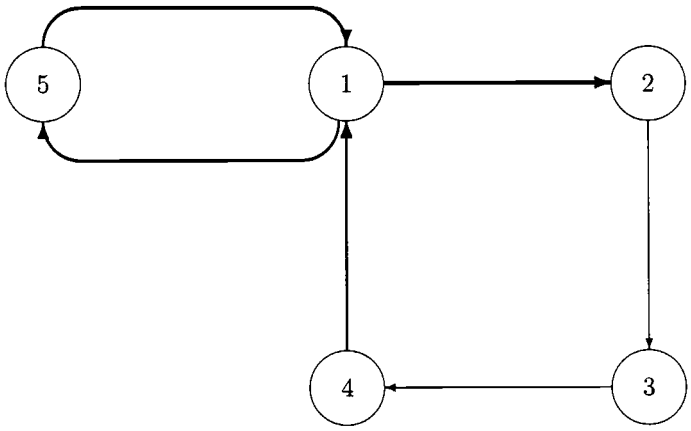


FIG. 4.

3. Periodic initial conditions and circulant matrices. In this section we turn our attention to some conditions on the pair (A, B) which ensure a strictly positive asymptotic dynamics for the associated 2D system (2), under the assumption that the initial conditions $\mathcal{X}_0$ have a periodic pattern.

Although this situation is admittedly restrictive, it deserves a thorough discussion for at least two reasons. First, it develops intuitive insights into the combinatorial and spectral properties of a positive matrix pair, meanwhile enlightening some interesting features of block circulant positive matrices. Second, this analysis leads the way to the solution of the general problem, which we shall afford in the subsequent section. If $\mathcal{X}_0$ is nonzero and periodic with period T, i.e.,

$$\mathbf{x}(\ell, -\ell) = \mathbf{x}(\ell + T, -\ell - T) \geq 0 \qquad \forall\ \ell \in \mathbb{Z}, \tag{20}$$

it is clear that the local states $\mathbf{x}(t + \ell, -\ell)$ on each subsequent separation set $\mathcal{C}_t$ still constitute a periodic sequence of period T. It is a matter of simple computation to check that the nT-dimensional vector

$$\mathbf{p}_T(t) := \begin{bmatrix} \mathbf{x}(t, 0) \\ \mathbf{x}(t + 1, -1) \\ \vdots \\ \mathbf{x}(t + T - 1, -T + 1) \end{bmatrix}, \tag{21}$$

obtained by stacking T consecutive local states on $\mathcal{C}_t$, updates according to the following equation:

$$\mathbf{p}_T(t + 1) = C_T(A, B)\ \mathbf{p}_T(t), \tag{22}$$

where $C_T(A, B)$ denotes the $nT \times nT$ block circulant matrix

$$C_T(A, B) = \begin{bmatrix} A & B & & & \\ & A & B & & \\ & & \ddots & \ddots & \\ & & & & B \\ B & & & & A \end{bmatrix} \tag{23}$$

if $T > 1$, and the $n \times n$ matrix $A + B$ if $T = 1$, i.e., if all initial local states on $\mathcal{C}_0$ coincide.

It is worth noticing that $\mathbf{p}_T(t)$ is completely determined by the initial condition, $\mathbf{p}_T(0) > 0$, and by the structure of $C_T(A, B)$. In particular,

• if $C_T(A, B)$ is irreducible, no component of $\mathbf{p}_T(t)$ remains permanently unexcited. Conversely, if $C_T(A, B)$ is reducible, a positive vector $\mathbf{p}_T(0)$ can be found such that for some $j \in \{1, 2, \ldots, nT\}$, the jth entry of $\mathbf{p}_T(t)$ is zero for all $t \in \mathbb{N}$.

• If $C_T(A, B)$ is irreducible, $\mathbf{p}_T(t)$ eventually becomes strictly positive if and only if the set of the indices corresponding to nonzero entries in $\mathbf{p}_T(0)$ includes at least one element of each communicating class in $\mathcal{D}(C_T(A, B))$.

• The matrix $C_T(A, B)$ is primitive if and only if for every $\mathbf{p}_T(0) > 0$ the vector $\mathbf{p}_T(t)$ eventually becomes strictly positive.

So, under the assumption of periodic initial conditions with period T, the asymptotic strict positivity of every state evolution of (2) is equivalent to the primitivity of $C_T(A, B)$, which describes the system dynamics according to (22). Consequently, our primary goal in this section is to investigate how the properties of a positive pair (A, B) affect those of $C_T(A, B)$ and, in particular, under what conditions $C_T(A, B)$ is irreducible or primitive. The solution of this problem relies on the results obtained in the previous section and on a couple of technical lemmas, available in the literature. The first lemma introduces a general result on the spectra of block circulant matrices, which allows to express the spectrum of $C_T(A, B)$ in terms of the spectra $\Lambda(A + e^{i2\pi\ell/T}B)$, $\ell = 0, 1, \ldots, T - 1$. The second lemma provides a useful criterion for recognizing irreducible matrices.

Lemma on circulant matrices (see [5]). The spectrum of the block circulant matrix

$$C = \begin{bmatrix} A_1 & A_2 & \ldots & A_T \\ A_T & A_1 & & A_{T-1} \\ & & \ddots & \\ A_2 & A_3 & \ldots & A_1 \end{bmatrix}, \qquad A_i \in \mathbb{R}^{n\times n},$$

is the nT-tuple given by

$$\begin{aligned}\Lambda(C) = \Lambda(A_1 + A_2 + \cdots + A_T) \uplus \Lambda(A_1 + e^{i\omega}A_2 + \cdots + e^{i\omega(T-1)}A_T) \\ \uplus \cdots \uplus \Lambda(A_1 + e^{i\omega(T-1)}A_2 + \cdots + e^{i\omega(T-1)(T-1)}A_T),\end{aligned}$$

where $\omega = 2\pi/T$. In particular, the spectrum of (23) is

$$\Lambda(C_T(A, B)) = \Lambda(A + B) \uplus \Lambda(A + e^{i\omega}B) \uplus \cdots \uplus \Lambda(A + e^{i\omega(T-1)}B). \tag{24}$$

Irreducibility criterion (see [15]). An $n \times n$ matrix $F > 0$ with a simple maximal eigenvalue $\lambda_{\max}$ is irreducible if and only if both F and F^T have strictly positive eigenvectors corresponding to $\lambda_{\max}$.

LEMMA 3.1. *Let $(A, B) \in \mathcal{I}_n$ and $T \in \mathbb{N}$. The circulant matrix $C_T(A, B)$ is*

i) *irreducible if and only if 1 is not an eigenvalue of any one of the following matrices:*

$$A + e^{i\omega}B, A + e^{2i\omega}B, \ldots, A + e^{(T-1)i\omega}B, \qquad \omega = 2\pi/T; \tag{25}$$

ii) *primitive if and only if $A + B$ is primitive and none of the above matrices has an eigenvalue of unitary modulus.*

Proof. i) By the above lemma on circulant matrices, if none of the matrices in (25) has 1 as an eigenvalue, 1 is the simple maximal eigenvalue of $A+B$ and hence of $C_T(A,B)$. On the other hand, if $\mathbf{v}$ and $\mathbf{w}$ denote two strictly positive eigenvectors of $A+B$ and $(A+B)^T$, respectively, corresponding to the eigenvalue 1, we have

$$C_T(A,B)\begin{bmatrix}\mathbf{v}\\ \vdots\\ \mathbf{v}\end{bmatrix}=\begin{bmatrix}\mathbf{v}\\ \vdots\\ \mathbf{v}\end{bmatrix},$$

and

$$C_T(A,B)^T\begin{bmatrix}\mathbf{w}\\ \vdots\\ \mathbf{w}\end{bmatrix}=\begin{bmatrix}\mathbf{w}\\ \vdots\\ \mathbf{w}\end{bmatrix}.$$

Consequently, both $C_T(A,B)$ and $C_T(A,B)^T$ have a strictly positive eigenvector corresponding to the eigenvalue 1, and hence are irreducible. Conversely, if 1 is an eigenvalue of some matrix in (25), the multiplicity of 1 as maximal eigenvalue of $C_T(A,B)$ is greater than one, and $C_T(A,B)$ is reducible.

ii) Assume that $C_T(A,B)$ is primitive. As its spectral radius $\rho(C_T(A,B))=1$ is an eigenvalue of $A+B$, none of the matrices $A+e^{i\omega\ell}B$, $\ell=1,2,\dots,T-1$, has an eigenvalue of unitary modulus. In particular, the irreducible matrix $A+B$, having no eigenvalue of unitary modulus except for 1, is primitive. Conversely, if $A+B$ is primitive and none of the matrices in (25) has an eigenvalue of unitary modulus, by the first part of the proof $C_T(A,B)$ is an irreducible matrix with 1 as simple maximal eigenvalue. As any other eigenvalue of $C_T(A,B)$ has modulus strictly less than 1, $C_T(A,B)$ must be primitive. □

It is easy to obtain dual statements for the block circulant matrices $C_T(B,A)$, $T=1,2,\dots$, thus relating the irreducibility and primitivity of these matrices to the spectra $\Lambda(e^{i\theta\ell}A+B)$, $\theta=2\pi/T$, $\ell=1,2,\dots,T-1$. In general, however, the irreducibility of $C_T(B,A)$ need not imply that of $C_T(A,B)$, as a consequence of the fact that the imprimitivity indices $h_A(B)$ and $h_B(A)$ need not coincide.

Example 5. The pair of matrices (A,B), with

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

is an element of $\mathcal{I}_3$. It is immediate to see from the digraphs $\mathcal{D}_A(B)$ and $\mathcal{D}_B(A)$ that the block circulant matrix $C_2(A,B)=\left[\begin{smallmatrix}A&B\\B&A\end{smallmatrix}\right]$ is reducible, whereas $C_2(B,A)=\left[\begin{smallmatrix}B&A\\A&B\end{smallmatrix}\right]$ is irreducible.

Notice that, different from the case of irreducibility, A and B play a symmetric role in determining the primitivity of $C_T(A,B)$. Actually, if $C_T(A,B)$ is primitive, none of the matrices $A+e^{i\omega\ell}B$, $\omega=2\pi/T$ and $\ell=1,2,\dots,T-1$, has an eigenvalue of unitary modulus, and this happens if and only if the same holds true for the family $B+e^{i\omega\ell}A$, $\ell=1,2,\dots,T-1$. Thus, $C_T(B,A)$ is primitive, too.

PROPOSITION 3.2. *Let $(A,B)\in\mathcal{I}_n$. The following facts are equivalent:*

i) *all circulant matrices $C_T(A,B)$, $T=1,2,\dots$, are irreducible*;

ii) $1\in\Lambda(A+e^{i\omega}B)$ *for some real number ω implies $\omega\equiv 0 \bmod 2\pi$.*

Proof. i) $\Rightarrow$ ii) Assume, by contradiction, that 1 is an eigenvalue of $A + e^{i\omega}B$, for some $\omega \not\equiv 0 \bmod 2\pi$. By Proposition 2.3, ω must be a rational multiple of 2π, i.e., $\omega = 2\pi(\nu/\bar{T})$, for some nonzero integers ν and $\bar{T}$, $\nu \not\equiv 0 \bmod \bar{T}$. But in this case, by Lemma 3.1, $C_{\bar{T}}(A, B)$ is a reducible matrix, which contradicts assumption i).

ii) $\Rightarrow$ i) This follows from Lemma 3.1, too. □

PROPOSITION 3.3. *Let $(A, B) \in \mathcal{I}_n$, with $A + B$ primitive. The following facts are equivalent:*

i) *all circulant matrices $C_T(A, B)$, $T = 1, 2, \ldots$, are primitive*;

ii) $1 \in \Lambda(e^{i\theta}A + e^{i\omega}B)$ *implies* $\theta \equiv \omega \equiv 0 \bmod 2\pi$.

Proof. i) $\Rightarrow$ ii) As remarked at the end of the previous section, $L_{A,B}$ has rank 2. So, all solutions of

$$L_{A,B} \begin{bmatrix} \theta \\ \omega \end{bmatrix} \equiv \mathbf{0} \bmod 2\pi$$

and, consequently, all the pairs (θ, ω) for which 1 belongs to $\Lambda(e^{i\theta}A + e^{i\omega}B)$ must be rational multiples of 2π, namely, $(\theta, \omega) = 2\pi(q_1, q_2)$, $q_1, q_2 \in \mathbb{Q}$. However, if 1 would be in $\Lambda(e^{i\theta}A + e^{i\omega}B)$, for certain θ and ω rational multiples of 2π, we would have $e^{-i\theta} \in \Lambda(A + e^{i(\omega-\theta)}B)$, thus contradicting Lemma 3.1.

ii) $\Rightarrow$ i) This is immediate from Lemma 3.1. □

Tying together Propositions 2.2 and 2.4 with the above results, several alternative characterizations of the irreducibility and the primitivity of all circulant matrices $C_T(A, B)$, $T \in \mathbb{N}$, can be obtained, based on the digraphs $\mathcal{D}_A(B)$ and $\mathcal{D}^*(A, B)$, respectively. In particular, graph-theoretic criteria are available for checking the above properties and hence the strict positivity of the asymptotic dynamics of (2), starting from periodic initial conditions.

4. Arbitrary initial conditions and 2D primitivity. In this section we drop the periodicity assumption and turn our attention to general (admissible) initial conditions. As every periodic $\mathcal{X}_0$ is admissible, it is clear that the primitivity of all $C_T(A, B)$, $T \in \mathbb{N}$, is necessary for 2D primitivity. We aim to prove that it is also sufficient.

In fact, we will show that when all $C_T(A, B)$, $T \in \mathbb{N}$, are primitive, and hence $L_{A,B}$ is right prime, there exists a solid convex cone $\mathcal{K}$ in $\mathbb{R}^2_+$ such that for all (h, k) in $\mathcal{K} \cap \mathbb{Z}^2$ the Hurwitz products $A^h \sqcup\!\sqcup^k B$ are strictly positive. Consequently, every nonzero local state $\mathbf{x}(\ell, -\ell) > 0$ produces a strictly positive state evolution inside the cone $(\ell, -\ell) + \mathcal{K}$, as we have

$$\mathbf{x}(h + \ell, k - \ell) \geq (A^h \sqcup\!\sqcup^k B)\mathbf{x}(\ell, -\ell) \gg 0 \qquad \forall\ (h + \ell, k - \ell) \in (\ell, -\ell) + \mathcal{K}.$$

The admissibility assumption on $\mathcal{X}_0$ guarantees that the union of all cones $(\ell, -\ell) + \mathcal{K}$, which correspond to positive initial states $\mathbf{x}(\ell, -\ell)$, includes all separation sets $\mathcal{C}_t$, for t greater than a suitable $t_{\min}$. Consequently, for $t > t_{\min}$ all local states on the separation set $\mathcal{C}_t$ are strictly positive.

The subsequent discussion is based on the following number theoretic result, which extends a well-known lemma attributed to Schur [3].

LEMMA 4.1. *Let $\mathcal{S}$ be a nonempty subset of $\mathbb{N}^2$, closed under addition, such that the $\mathbb{Z}$-module generated by $\mathcal{S}$ is $\mathbb{Z}^2$. Then there exists a solid convex cone $\mathcal{K}^*$ in $\mathbb{R}^2_+$ such that all elements in $\mathcal{K}^* \cap \mathbb{Z}^2$ are in $\mathcal{S}$.*

Proof. Let (α_1, β_1), …, (α_t, β_t) be a set of elements of $\mathcal{S}$ which generate $\mathbb{Z}^2$, and let $r := \sum_{i=1}^{t}(\alpha_i + \beta_i)$. For every nonnegative pair (h, k) in $\mathcal{T} := \{(h, k) : h, k \in$

$\mathbb{N}, h+k \le r\}$ we may determine integer coefficients $c_i^{h,k}$ such that

$$(h,k) = \sum_{i=1}^{t} c_i^{h,k}(\alpha_i, \beta_i).$$

Let M be the maximum of the integers $|c_i^{h,k}|$, $(h,k) \in \mathcal{T}$, and $i = 1, 2, \ldots, t$, and define

$$(v,w) := \sum_{i=1}^{t} M\ (\alpha_i, \beta_i).$$

As the $\mathbb{Z}$-module generated by $\mathcal{S}$ is $\mathbb{Z}^2$, the cone $\mathcal{K}$ generated in $\mathbb{R}_+^2$ by the positive pairs (α_1, β_1), …, (α_t, β_t) is convex and solid. We aim to show that all integer pairs in $\mathcal{K}^* := (v,w) + \mathcal{K}$ belong to $\mathcal{S}$. Every integer pair (c,d) in $\mathcal{K}$ can be expressed as

$$(c,d) = \sum_{i=1}^{t} q_i\ (\alpha_i, \beta_i), \qquad q_i \in \mathbb{Q}_+,$$

and therefore as

$$(c,d) = \sum_{i=1}^{t} \lfloor q_i \rfloor (\alpha_i, \beta_i) + \sum_{i=1}^{t} (q_i - \lfloor q_i \rfloor)(\alpha_i, \beta_i),$$

where $\lfloor q_i \rfloor$ denotes the integer part of q_i. Since $0 \le q_i - \lfloor q_i \rfloor < 1$, the pair $(\bar{c}, \bar{d}) := \sum_{i=1}^{t} (q_i - \lfloor q_i \rfloor)(\alpha_i, \beta_i)$ is an element of $\mathcal{T}$, and (c,d) decomposes into

$$(c,d) = (\bar{c}, \bar{d}) + \sum_{i=1}^{t} n_i\ (\alpha_i, \beta_i), \qquad n_i \in \mathbb{N}. \tag{26}$$

So, every integer pair (h,k) in $\mathcal{K}^*$ can be written as $(h,k) = (v,w) + (c,d)$, $(c,d) \in \mathcal{K}$, and hence as

$$\begin{aligned}(h,k) &= (v,w) + (\bar{c}, \bar{d}) + \sum_{i=1}^{t} n_i\ (\alpha_i, \beta_i) \\ &= \sum_{i=1}^{t} M\ (\alpha_i, \beta_i) + \sum_{i=1}^{t} c_i^{\bar{c},\bar{d}}(\alpha_i, \beta_i) + \sum_{i=1}^{t} n_i\ (\alpha_i, \beta_i) \\ &= \sum_{i=1}^{t} (M + n_i + c_i^{\bar{c},\bar{d}})(\alpha_i, \beta_i),\end{aligned}$$

with n_i and $c_i^{\bar{c},\bar{d}}$ in $\mathbb{N}$, $i = 1, 2, \ldots, t$. Since $M + c_i^{\bar{c},\bar{d}} + n_i$ is a nonnegative integer for every i, and $\mathcal{S}$ is closed under addition, (h,k) belongs to $\mathcal{S}$. □

PROPOSITION 4.2. *Let (A,B) be in $\mathcal{I}_n$. The following facts are equivalent:*

i) *the integer matrix $L_{A,B}$ is right prime*;

ii) *there is a solid convex cone $\mathcal{K}$ in $\mathbb{R}_+^2$ such that for every pair of integers (h,k) in $\mathcal{K}$ and every couple of vertices i and j, there is a path p in $\mathcal{D}^*(A,B)$, from i to j, including h A-arcs and k B-arcs*;

iii) *there is a solid convex cone $\mathcal{K}_H$ in $\mathbb{R}_+^2$ such that for every pair of integers (h,k) in $\mathcal{K}_H$ the Hurwitz product $A^h \sqcup\!\sqcup^k B$ is strictly positive*;

iv) *the pair* (A, B) *is primitive.*

Proof. i) $\Rightarrow$ ii) Let $\mathcal{S}_\ell$ be the set of integer vectors $[\alpha(\gamma)\ \beta(\gamma)]$ corresponding to all cycles γ in $\mathcal{D}^*(A, B)$ passing through vertex ℓ. Clearly, $\mathcal{S}_\ell$ is nonempty and closed under addition. Moreover, the $\mathbb{Z}$-module generated by $\mathcal{S}_\ell$ coincides with the $\mathbb{Z}$-module generated by the rows of $L_{A,B}$, namely, with $\mathbb{Z}^2$. Actually, consider a positive vector $[\alpha(\gamma)\ \beta(\gamma)]$, γ a circuit in $\mathcal{D}^*(A, B)$, which is not included in $\mathcal{S}_\ell$, and let j be any vertex γ passes through. As $\mathcal{D}^*(A, B)$ is strongly connected, it includes a cycle γ' passing through ℓ and j, and another cycle, γ'', obtained by connecting γ and γ'. So, both $[\alpha(\gamma')\ \beta(\gamma')]$ and $[\alpha(\gamma'')\ \beta(\gamma'')]$ are in $\mathcal{S}_\ell$, and

$$[\alpha(\gamma)\ \beta(\gamma)] = [\alpha(\gamma'')\ \beta(\gamma'')] - [\alpha(\gamma')\ \beta(\gamma')]$$

is in the $\mathbb{Z}$-module generated by $\mathcal{S}_\ell$. By the above lemma, then, there exists a solid convex cone $\mathcal{K}_\ell^*$ in $\mathbb{R}_+^2$ such that all integer vectors in $\mathcal{K}_\ell^*$ are in $\mathcal{S}_\ell$.

If i and j are arbitrary vertices in $\mathcal{D}^*(A, B)$ and $p_{i\ell}$ and $p_{\ell j}$ are two fixed paths connecting i to ℓ and ℓ to j, respectively, all integer vectors in the cone

$$\mathcal{K}_{ij}^* := [\alpha(p_{i\ell}) + \alpha(p_{\ell j}) \quad \beta(p_{i\ell}) + \beta(p_{\ell j})] + \mathcal{K}_\ell^*$$

correspond to paths connecting i to j. Clearly, $\mathcal{K} := \cap_{ij}\mathcal{K}_{ij}^*$ is a solid convex cone which satisfies ii).

ii) $\Rightarrow$ iii) This is obvious, once $\mathcal{K}_H = \mathcal{K}$ is assumed.

iii) $\Rightarrow$ iv) Under assumption iii), it is easy to see that every admissible $\mathcal{X}_0$ eventually produces a strictly positive state evolution, and hence the pair (A, B) is primitive, by definition.

iv) $\Rightarrow$ i) When (A, B) is primitive, all nonzero periodic initial conditions eventually produce strictly positive dynamics. This implies that all $C_T(A, B)$, $T \in \mathbb{N}$, are primitive matrices and hence, by Propositions 2.2 and 3.3, $L_{A,B}$ is right prime. ∎

To conclude, observe that the above proposition reduces the primitivity of the pair (A, B) to the existence of a solid cone $\mathcal{K}_H$ in $\mathbb{R}_+^2$, whose integer coordinates points correspond to strictly positive Hurwitz products. Indeed, this condition can be considerably simplified, as the existence of a primitive, and hence of a strictly positive, Hurwitz product ensures that of a whole cone $\mathcal{K}_H$ of strictly positive Hurwitz products. This property nicely extends to matrix pairs the well-known fact that a positive matrix F is primitive if and only if it has a strictly positive power.

PROPOSITION 4.3. *Let* (A, B) *be in* $\mathcal{I}_n$. *The following facts are equivalent:*

i) *there is a solid convex cone* $\mathcal{K}_H$ *in* $\mathbb{R}_+^2$ *such that for every pair of integers* (h, k) *in* $\mathcal{K}_H$ *the Hurwitz product* $A^h \sqcup\!\sqcup^k B$ *is strictly positive*;

ii) *there exists a positive pair* $(\ell, m) \in \mathbb{N} \times \mathbb{N}$ *such that* $A^\ell \sqcup\!\sqcup^m B$ *is primitive.*

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

ii) $\Rightarrow$ i) Assume that $A^\ell \sqcup\!\sqcup^m B$ is primitive. Then there exists a positive integer r such that $A^{r\ell} \sqcup\!\sqcup^{rm} B \geq (A^\ell \sqcup\!\sqcup^m B)^r \gg 0$. So, it is not restrictive to assume that $A^\ell \sqcup\!\sqcup^m B$ is strictly positive. As A and B are both positive and $\mathcal{D}^*(A, B)$ is strongly connected, there exists a vertex j with an outgoing A-arc, (j, u), and an ingoing B-arc (e, j). By the assumption on $A^\ell \sqcup\!\sqcup^m B$, in $\mathcal{D}^*(A, B)$ one can find a cycle γ passing through j, a path p_{uj} from u to j and a path p_{je} from j to e, each of them including ℓ A-arcs and m B-arcs. So, the path p_{ju} can be completed into a cycle with $\ell+1$ A-arcs and m B-arcs, and similarly p_{ej} can be completed into a cycle including ℓ A-arcs and $m+1$ B-arcs. Clearly, the $\mathbb{Z}$-module $\mathcal{M}$ generated by the pairs (ℓ, m), $(\ell+1, m)$, and

$(\ell, m+1)$ is $\mathbb{Z}^2$, as the integer matrix

$$\begin{bmatrix} \ell & m \\ \ell+1 & m \\ \ell & m+1 \end{bmatrix}$$

is right prime. Moreover, as the $\mathbb{Z}$-module generated by the rows of $L_{A,B}$ includes $\mathcal{M}$, $L_{A,B}$ is a right prime matrix, too, and the conclusion follows from the above proposition. □

REFERENCES

[1] B. D. O. Anderson, M. Deistler, L. Farina, and L. Benvenuti, *Nonnegative realization of a linear system with nonnegative impulse response*, IEEE Trans. Circuits and Systems, Part I, 43 (1996), pp. 134–142.

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

[3] R. A. Brualdi and H. J. Ryser, *Combinatorial Matrix Theory*, Cambridge University Press, Cambridge, UK, 1991.

[4] P. G. Coxson and H. Shapiro, *Positive reachability and controllability of positive systems*, Linear Algebra Appl., 94 (1987), pp. 35–53.

[5] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1979.

[6] M. P. Fanti, B. Maione, and B. Turchiano, *Controllability of linear single-input positive discrete time systems*, Internat. J. Control, 50 (1989), pp. 2523–2542.

[7] M. P. Fanti, B. Maione, and B. Turchiano, *Controllability of multi-input positive discrete time systems*, Internat. J. Control, 51 (1990), pp. 1295–1308.

[8] E. Fornasini and G. Marchesini, *Doubly indexed dynamical systems*, Math. Systems Theory, 12 (1978), pp. 59–72.

[9] E. Fornasini, *A 2D systems approach to river pollution modelling*, Multidimens. Systems Signal Process., 2 (1991), pp. 233–265.

[10] E. Fornasini and M. E. Valcher, *Matrix pairs in 2D systems: An approach based on trace series and Hankel matrices*, SIAM J. Control Optim., 33 (1995), pp. 1127–1150.

[11] E. Fornasini and M. E. Valcher, *On the spectral and combinatorial structure of 2D positive systems*, Linear Algebra Appl., 245 (1996), pp. 223–258.

[12] W. P. Heath, *Self-Tuning Control for Two-Dimensional Processes*, John Wiley, New York, 1994.

[13] N. Jacobson, *Basic Algebra*, Vols. I and II, W. H. Freeman, San Francisco, CA, 1974.

[14] H. Maeda and S. Kodama, *Positive realization of difference equations*, IEEE Trans. Circuits and Systems, CAS-28 (1981), pp. 39–47.

[15] H. Minc, *Nonnegative Matrices*, John Wiley, New York, 1988.

[16] M. Newman, *Integral Matrices*, Academic Press, New York, 1972.

[17] M. E. Valcher, *Controllability and reachability criteria for discrete time positive systems*, Internat. J. Control, 65 (1996), pp. 511–536.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 89–98, January 1998

006

BRUHAT DECOMPOSITION AND NUMERICAL STABILITY*

O. H. ODEH†, D. D. OLESKY‡, AND P. VAN DEN DRIESSCHE§

Abstract. For a real nonsingular n-by-n matrix A, there exists a decomposition $A = V\Pi U$, where Π is a permutation matrix and V, U are upper triangular matrices. When $\Pi^T V \Pi$ is lower triangular and U is normalized, such a decomposition is called the left Bruhat decomposition of A. An algorithm for computing the left Bruhat decomposition is given. For classes of matrices introduced by Wilkinson and recently (from a practical application) by Foster that have an exponential growth factor when Gaussian elimination with partial pivoting (GEPP) is applied, left Bruhat decomposition has at most linear growth. A partial pivoting strategy for Bruhat decomposition is also developed, and an explicit equivalence between GEPP and Bruhat decomposition with partial pivoting (BDPP) is derived. This equivalence implies that the growth factor for GEPP on A equals the growth factor for BDPP on ρA^T, where ρ is the permutation matrix that reverses the rows of A^T. BDPP is shown to give a growth factor of at most 2 when applied to any matrix for which GEPP gives the maximal growth factor of 2^{n-1}.

Key words. Bruhat decomposition, Gaussian elimination, growth factor, numerical stability, partial pivoting

AMS subject classifications. 65F05, 15A23

PII. S0895479896303314

1. Introduction.

Matrix factorization techniques are frequently used for solving nonsingular systems of linear equations. The most common factorization is LU decomposition, and Gaussian elimination with partial pivoting (GEPP) is the most common practical algorithm for computing an LU decomposition. However, other decompositions, such as LPR decomposition (see, e.g., Elsner [2], Gohberg and Goldberg [4]) and Bruhat decomposition, can also be used to solve linear systems.

Bruhat decomposition, known from the theory of linear algebraic groups [5], [9], was considered by Kolotilina and Yeremin [9] as an alternative to LU decomposition for solving sparse systems of linear equations. Kolotilina and Yeremin also gave relations between Bruhat decomposition and the other two decompositions given above, and sparsity of the Bruhat decomposition factors was considered in [8].

In the following sections we describe the left Bruhat decomposition, and give an algorithm for its computation (Algorithm 2.1), which is the analogue of an algorithm given in [9, section 2] for the right Bruhat decomposition. In contrast with GEPP, Bruhat decomposition is numerically stable for the classes of matrices given by Wilkinson and Foster (section 3). We also introduce a pivoting strategy for Bruhat decomposition (Algorithm 4.1) and derive explicit relationships (Corollary 4.5) between the factors that are determined by applying GEPP to A and Bruhat decomposition with partial pivoting (BDPP) to ρA^T, where ρ is the permutation matrix that re-

*Received by the editors May 10, 1996; accepted for publication (in revised form) by N. J. Higham December 20, 1996. The research of the second and third authors was partially supported by the Natural Science and Engineering Research Council of Canada and the University of Victoria Committee on Faculty Research and Travel.

http://www.siam.org/journals/simax/19-1/30331.html

†The author is deceased. Former address: Department of Computer Science, University of Victoria, Victoria, BC, Canada V8W 3P6.

‡Department of Computer Science, University of Victoria, Victoria, BC, Canada V8W 3P6 (dolesky@csr.uvic.ca).

§Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada V8W 3P4 (pvdd@smart.math.uvic.ca).

verses the order of the rows of A^T. We show that BDPP gives a growth factor of at most 2 when applied to matrices that give maximal growth when GEPP is applied (section 5). BDPP is a practical algorithm for solving systems of linear equations and is an alternative to consider when GEPP may be unstable.

2. Description of the Bruhat decomposition. Let A be a given n-by-n real nonsingular matrix. Then there exists a decomposition

$$A = V\Pi U, \tag{2.1}$$

where V and U are n-by-n upper triangular matrices and Π is an n-by-n permutation matrix. The permutation matrix Π in (2.1) is uniquely determined by A [9]. A decomposition of the form (2.1) is called a *Bruhat decomposition* of the matrix A, and Π is called the *Bruhat permutation of* A. The decomposition (2.1) is called the *reduced on the left* Bruhat decomposition if the matrix $\Pi^T V \Pi$ is lower triangular, and *reduced on the right* if the matrix $\Pi U \Pi^T$ is lower triangular [9]. For the remainder of this paper, we work with the reduced on the left Bruhat decomposition with U normalized to have all diagonal entries equal to 1, and we refer to this as the *left Bruhat decomposition*. With this normalization, the left Bruhat decomposition of a given nonsingular matrix is unique.

The decomposition (2.1) can be computed by postmultiplication of A by $n-1$ nonsingular matrices $U^{(i)}$, whose entries are chosen so as to introduce zeros into the matrix product. Let

$$A^{(0)} = A \quad \text{and} \quad A^{(i)} = A^{(i-1)}U^{(i)}, \quad 1 \le i \le n-1,$$

so that $A^{(i)} = AU^{(1)}U^{(2)}\cdots U^{(i)}$.

Denoting $A^{(i)} = [a_{jk}^{(i)}]$, the matrices $U^{(i)}$ can be written compactly as

$$U^{(i)} = I - e^{(i)}\left(m^{(i)}\right)^T,$$

where

$$m_j^{(i)} = \begin{cases} \frac{a_{r_i,j}^{(i-1)}}{a_{r_i,i}^{(i-1)}} & \text{for } i+1 \le j \le n, \\ 0 & \text{otherwise,} \end{cases}$$

$$e_j^{(i)} = \begin{cases} 1, & i = j, \\ 0 & \text{otherwise,} \end{cases}$$

and r_i is the maximum row index such that $a_{r_i,i}^{(i-1)} \neq 0$. Thus, at the ith step, $a_{r_i,i}^{(i-1)}$ is the pivot entry, and multiplication by $U^{(i)}$ zeros out all entries of $A^{(i-1)}$ in row r_i and columns $i+1,\ldots,n$. After $n-1$ elimination steps, $A^{(n-1)} = AU^{(1)}U^{(2)}\cdots U^{(n-1)}$. Let $a_{r_n,n}^{(n-1)}$ denote the sole nonzero entry in column n of $A^{(n-1)}$ and $\Pi = [\pi_{jk}]$ be the permutation matrix with $\pi_{r_k,k} = 1$ for $1 \le k \le n$. Then, letting

$$V = A^{(n-1)}\,\Pi^T \tag{2.2}$$

and $U^{-1} = U^{(1)}U^{(2)}\cdots U^{(n-1)}$ gives $A = V\Pi U$.

The following algorithm determines the factors of this decomposition.

ALGORITHM 2.1 (left Bruhat decomposition).

Input: Nonsingular n-by-n matrix A

Output: The matrices V, Π, and U, where the left Bruhat decomposition is $A = V\Pi U$

Initialization: $U = I$

for $i = 1$ *to* n
  $j = \max\{p \mid a_{pi} \neq 0\}$
  $\pi_{ji} = 1, \quad \pi_{\ell i} = 0$ *for* $\ell \neq j$
  $\nu_{tj} = a_{ti}$ *for* $1 \leq t \leq n$
  for $k = i + 1$ *to* n
    $m = \dfrac{a_{jk}}{a_{ji}}$
    $u_{ik} = m$
    for $\ell = 1$ *to* $j - 1$
      $a_{\ell k} = a_{\ell k} - m a_{\ell i}$
    $a_{jk} = 0$

By construction, U is upper triangular. Thus, to prove that Algorithm 2.1 gives the left Bruhat decomposition of A, we show that V is upper triangular, and then we show that $\Pi^T V \Pi$ is lower triangular. Let $\pi(j) = i$ if $\pi_{ji} = 1$; then $\pi^{-1}(i) = j$. From (2.2), for any fixed q, $\nu_{iq} = a_{i,\pi(q)}^{(n-1)}$. If $q = \max\{p \mid a_{p,\pi(q)}^{(n-1)} \neq 0\}$, then $a_{i,\pi(q)}^{(n-1)} = 0$ for $i > q$, hence V is upper triangular. Also by (2.2)

$$(\Pi^T V \Pi)_{rj} = (\Pi^T A^{(n-1)})_{rj} = a_{\pi^{-1}(r),j}^{(n-1)}.$$

But $a_{\pi^{-1}(r),r}^{(n-1)} \neq 0$ and $a_{\pi^{-1}(r),j}^{(n-1)} = 0$ for $j > r$ as these entries are eliminated in the rth step of the algorithm. Hence, $\Pi^T V \Pi$ is lower triangular.

In general, Π cannot be determined from the zero–nonzero pattern of A; it depends as well on the numerics. Even if matrix A does not have an LU decomposition, there exists a permutation matrix P such that PA has an LU decomposition. Such a permutation matrix is Π^T from the left Bruhat decomposition [9], because if $A = V\Pi U$, then $\Pi^T A = (\Pi^T V \Pi)U = LU$. This relationship between the left Bruhat decomposition of A and the LU decomposition (with U normalized) of $\Pi^T A$ shows that each of the triangular factors of the left Bruhat decomposition is uniquely determined.

3. Bruhat decomposition of matrices with large γ for GEPP. For GEPP on a nonsingular matrix $A = [a_{jk}]$, the growth factor γ is defined as

$$\gamma = \max_{i,j,k} |a_{jk}^{(i)}| \;/\; \max_{j,k} |a_{jk}|,$$

where $A^{(i)} = [a_{jk}^{(i)}]$ is the derived matrix after the ith elimination step (see, e.g., [7, p. 177] and [10, p. 151]). The computation of the solution x of a linear system $Ax = b$ may be unstable if the growth factor is very large [6]. Motivated by a backward error analysis for LU decomposition [7, p. 176] and its relationship to Bruhat decomposition, we define the *growth factor* for Bruhat decomposition as

$$\gamma_B = \max\left\{ \max_{i,j,k} |u_{jk}^{(i)}| / \max_{j,k} |a_{jk}|,\ \max_{i,j,k} |a_{jk}^{(i)}| / \max_{j,k} |a_{jk}| \right\}. \tag{3.1}$$

Wilkinson [11, p. 212] introduced an n-by-n matrix W_n that achieves the largest possible growth factor of 2^{n-1} when GEPP is applied. The Bruhat decomposition, on the other hand, gives $\gamma_B = 2$, as demonstrated in the following example.

Example 3.1. The left Bruhat decomposition of the 5-by-5 Wilkinson matrix is

$$W_5 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -1 & -\frac{1}{2} & -\frac{1}{4} & 1 \\ 0 & 2 & 0 & 0 & -1 \\ 0 & 0 & 2 & 0 & -1 \\ 0 & 0 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & -1 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 1 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

In general, application of Algorithm 2.1 to the n-by-n Wilkinson matrix W_n gives $\gamma_B = 2$.

Wilkinson also noted that matrices with large γ do not seem to arise in practical applications. However, recently, Foster [3] discussed a class of n-by-n matrices that arises in the numerical solution of Volterra integral equations and that for GEPP has growth factor close to the maximal value of 2^{n-1}. In contrast, for Bruhat decomposition on an n-by-n matrix in Foster's class, the factors V and U can be explicitly determined, and γ_B is linear in n.

Bruhat decomposition is a good alternative to GEPP for the matrices above when the latter gives exponentially large growth factors. However, for some matrices both GEPP and Bruhat decomposition give exponential growth (for example, the block matrix given by Wright [12, equations (10) and (12)]). There are also examples of matrices for which Bruhat decomposition gives exponential growth, whereas GEPP gives constant growth; one such example is ρW_n, where ρ is the permutation matrix that reverses the rows of W_n.

4. A pivoting strategy for Bruhat decomposition. We now present a pivoting strategy for Bruhat decomposition that, like the use of partial pivoting with Gaussian elimination, keeps the multipliers bounded by one and usually results in a stable computation. The decomposition is computed by postmultiplication of A by $n-1$ pairs of nonsingular matrices $P^{(i)}U^{(i)}$ for $i = 1, 2, \ldots, n-1$, where $P^{(i)}$ is a permutation matrix and $U^{(i)}$ is chosen to introduce zeros into the matrix product. Let $A^{(0)} = A$ and $A^{(i)} = A^{(i-1)}P^{(i)}U^{(i)}$, so that

$$A^{(i)} = AP^{(1)}U^{(1)}P^{(2)}U^{(2)} \cdots P^{(i)}U^{(i)}.$$

At the ith step of the decomposition, $P^{(i)}$ is chosen to interchange columns i and c of $A^{(i-1)}$, where c is such that

$$\max_{i \le t \le n} |a^{(i-1)}_{n-i+1,t}| = |a^{(i-1)}_{n-i+1,c}|.$$

Then $U^{(i)}$ is chosen so that $a^{(i)}_{n-i+1,r} = 0$ for $r = i+1, i+2, \ldots, n$. That is, letting $A^{(i-1)}P^{(i)} = [\tilde{a}^{(i-1)}_{jk}]$, then $U^{(i)} = I - e^{(i)}(m^{(i)})^T$, where

$$m^{(i)}_j = \begin{cases} \frac{\tilde{a}^{(i-1)}_{n-i+1,j}}{\tilde{a}^{(i-1)}_{n-i+1,i}} & \text{for } i+1 \le j \le n, \\ 0 & \text{otherwise.} \end{cases}$$

After $n-1$ steps,

$$\begin{aligned} A^{(n-1)} &= AP^{(1)}U^{(1)}P^{(2)}U^{(2)}\cdots P^{(n-1)}U^{(n-1)} \\ &= \begin{bmatrix} a^{(1)}_{11} & a^{(2)}_{12} & \cdots & a^{(n-2)}_{1,n-2} & a^{(n-1)}_{1,n-1} & a^{(n-1)}_{1n} \\ a^{(1)}_{21} & a^{(2)}_{22} & \cdots & a^{(n-2)}_{2,n-2} & a^{(n-1)}_{2,n-1} & \\ a^{(1)}_{31} & a^{(2)}_{32} & \cdots & a^{(n-2)}_{3,n-2} & & \\ \vdots & \iddots & & 0 & & \\ a^{(1)}_{n1} & & & & & \end{bmatrix} \\ &= V\rho, \end{aligned}$$

where V is an upper triangular matrix and the permutation matrix ρ reverses the columns of V.

The following algorithm essentially determines the factors of the above decomposition

$$A = V\rho\left(U^{(n-1)}\right)^{-1} P^{(n-1)}\left(U^{(n-2)}\right)^{-1} P^{(n-2)}\cdots\left(U^{(1)}\right)^{-1} P^{(1)},$$

where we note that $\left(P^{(i)}\right)^{-1} = P^{(i)}$ for $i = 1, 2, \ldots, n-1$. The one-dimensional array P has $P(i) = c$ if $P^{(i)}$ interchanges columns i and c of $A^{(i-1)}$. The ith row of the upper triangular matrix $\left(U^{(i)}\right)^{-1}$ is stored in the ith row of an n-by-n matrix U. The reduced matrices $A^{(i)}$ overwrite A and the function $swap(i, c)$ is used to interchange columns i and c of A.

ALGORITHM 4.1 (Bruhat decomposition with partial pivoting (BDPP)).

Input: Nonsingular n-by-n matrix A

Output: The essential components of the factors V, $\left(U^{(i)}\right)^{-1}$, and $P^{(i)}$ of the Bruhat decomposition with partial pivoting of A.

Initialization: $U = I, P(j) = j$ *for* $j = 1, 2, \ldots, n-1$

for $j = n$ *to* 2
  $i = n - j + 1$
  find $c : \max_{i \le t \le n} |a_{jt}| = |a_{jc}|$
  if $c > i$ *then*
    swap (i, c)
    $P(i) = c$
  $\nu_{tj} = a_{ti}$ *for* $1 \le t \le j$
  for $k = i+1$ *to* n
    $m = \dfrac{a_{jk}}{a_{ji}}$

$u_{ik} = m$
for $\ell = 1$ *to* $j-1$
$a_{\ell k} = a_{\ell k} - m a_{\ell i}$
$a_{jk} = 0$
$\nu_{11} = a_{1n}$

The next theorem shows an equivalence between BDPP and GEPP.

THEOREM 4.2. *Let A be an n-by-n nonsingular matrix. Suppose that*

$$L^{(n-1)}\bar{P}^{(n-1)}L^{(n-2)}\bar{P}^{(n-2)}\cdots L^{(1)}\bar{P}^{(1)}A = \bar{U}$$

is the result of applying GEPP *to A, where $L^{(i)}$ is the lower triangular matrix of multipliers and $\bar{P}^{(i)}$ is the permutation matrix associated with the ith step of GEPP. Suppose also that*

$$\rho A^T P^{(1)}U^{(1)}P^{(2)}U^{(2)}\cdots P^{(n-1)}U^{(n-1)} = V\rho$$

is the result of applying BDPP *to $\rho A^T = B$. Then $P^{(i)} = \bar{P}^{(i)}$, $U^{(i)} = (L^{(i)})^T$, and $\rho(A^{(i)})^T = B^{(i)}$ for $1 \le i \le n-1$.*

Proof. The proof is by induction. For $i = 1$, consider the first step of GEPP. Let

$$\max_{1\le j\le n} |a_{j1}| = |a_{t1}|.$$

Thus, the effect of $\bar{P}^{(1)}$ is to interchange rows 1 and t. Letting $\bar{P}^{(1)}A = [\tilde{a}_{jk}]$, then $L^{(1)} = I - m^{(1)}\left(e^{(1)}\right)^T$, where $m_j^{(1)} = \frac{\tilde{a}_{j1}}{\tilde{a}_{11}}$ is the jth entry of the vector $m^{(1)}$ for $2 \le j \le n$ and $m_1^{(1)} = 0$. Thus, $A^{(1)} = L^{(1)}P^{(1)}A$. Now consider the first step of BDPP applied to $B = \rho A^T$. Note that

$$b_{n-p+1,j} = \sum_{k=1}^{n} \rho_{n-p+1,k}\, a_{jk} = a_{jp} \quad \text{for} \quad 1 \le j,\ \ p \le n.$$

Thus,

$$\max_{1\le j\le n} |b_{nj}| = \max_{1\le j\le n} |a_{j1}| = |a_{t1}| = |b_{nt}|,$$

and the effect of $P^{(1)}$ is to interchange columns 1 and t, so that $P^{(1)} = \left(\bar{P}^{(1)}\right)^T = \bar{P}^{(1)}$. Let $BP^{(1)} = [\tilde{b}_{jk}]$, and note that $BP^{(1)} = \rho\left(\bar{P}^{(1)}A\right)^T$. Now $U^{(1)} = I - e^{(1)}\left(x^{(1)}\right)^T$, where

$$x_j^{(1)} = \frac{\tilde{b}_{nj}}{\tilde{b}_{n1}} = \frac{\tilde{a}_{j1}}{\tilde{a}_{11}} = m_j^{(1)} \quad \text{for} \quad 2 \le j \le n \quad \text{and} \quad x_1^{(1)} = m_1^{(1)} = 0.$$

Thus, $U^{(1)} = \left(L^{(1)}\right)^T$; hence,

$$B^{(1)} = BP^{(1)}U^{(1)} = \rho\left(L^{(1)}\,\bar{P}^{(1)}\,A\right)^T = \rho\left(A^{(1)}\right)^T.$$

Thus, the statement is true for $i = 1$.

Suppose that the theorem is true for all i such that $1 \le i \le s < n-1$, and consider the $(s+1)$st step of GEPP. Let

$$\max_{s+1 \le j \le n} |a_{j,s+1}^{(s)}| = |a_{r,s+1}^{(s)}|.$$

Thus, the effect of $\bar{P}^{(s+1)}$ is to interchange rows $(s+1)$ and r. Letting

$$\bar{P}^{(s+1)} A^{(s)} = \left[\tilde{a}_{jk}^{(s)}\right],$$

then $L^{(s+1)} = I - m^{(s+1)} \left(e^{(s+1)}\right)^T$, where

$$m_j^{(s+1)} = \frac{\tilde{a}_{j,s+1}^{(s)}}{\tilde{a}_{s+1,s+1}^{(s)}} \quad \text{for} \quad s+2 \le j \le n$$

and

$$m_j^{(s+1)} = 0 \quad \text{for} \quad 1 \le j \le s+1.$$

Thus, $A^{(s+1)} = L^{(s+1)} \bar{P}^{(s+1)} A^{(s)}$. Now consider the $(s+1)$st step of BDPP applied to $B = \rho A^T$. By the induction hypothesis, $B^{(s)} = \rho \left(A^{(s)}\right)^T$, and consequently

$$\max_{s+1 \le j \le n} \left|b_{n-s,j}^{(s)}\right| = \max_{s+1 \le j \le n} \left|a_{j,s+1}^{(s)}\right| = \left|a_{r,s+1}^{(s)}\right| = \left|b_{n-s,r}^{(s)}\right|.$$

Thus, the effect of $P^{(s+1)}$ is to interchange columns $(s+1)$ and r, so that

$$P^{(s+1)} = \left(\bar{P}^{(s+1)}\right)^T = \bar{P}^{(s+1)}.$$

Let $B^{(s)} P^{(s+1)} = [\tilde{b}_{jk}^{(s)}]$, and note that $B^{(s)} P^{(s+1)} = \rho \left(\bar{P}^{(s+1)} A^{(s)}\right)^T$. Now

$$U^{(s+1)} = I - e^{(s+1)} \left(x^{(s+1)}\right)^T,$$

where

$$x_j^{(s+1)} = \frac{\tilde{b}_{n-s,j}^{(s)}}{\tilde{b}_{n-s,s+1}^{(s)}} = \frac{\tilde{a}_{j,s+1}^{(s)}}{\tilde{a}_{s+1,s+1}^{(s)}} = m_j^{(s+1)} \quad \text{for} \quad s+2 \le j \le n$$

and

$$x_j^{(s+1)} = m_j^{(s+1)} = 0 \quad \text{for} \quad 1 \le j \le s+1.$$

Thus, $U^{(s+1)} = \left(L^{(s+1)}\right)^T$, and

$$B^{(s+1)} = BP^{(s+1)}U^{(s+1)} = \rho \left(L^{(s+1)} \bar{P}^{(s+1)} A\right)^T = \rho \left(A^{(s+1)}\right)^T,$$

completing the proof. □

Remark 4.3. Consider GEPP applied to $A^T\rho$. From Theorem 4.2, this is equivalent to the application of BDPP to $\rho(A^T\rho)^T = A$.

An immediate consequence of Theorem 4.2 is the following, which shows that Bruhat decomposition with partial pivoting on A determines the left Bruhat decomposition of a column permutation of A. This result is analogous to a well-known result for GEPP.

COROLLARY 4.4. *Suppose A is an n-by-n nonsingular matrix and let*

$$AP^{(1)}U^{(1)}P^{(2)}U^{(2)}\cdots P^{(n-1)}U^{(n-1)} = V\rho$$

be the result of BDPP *applied to A. Then there exist a permutation matrix P and an upper triangular matrix U such that $AP = V\rho U$.*

Proof. Let

$$L^{(n-1)}\,\bar{P}^{(n-1)}\,L^{(n-2)}\,\bar{P}^{(n-2)}\cdots L^{(1)}\,\bar{P}^{(1)}\,A^T\rho = L^{-1}\,\bar{P}\,A^T\rho = \bar{U}$$

be the result of applying GEPP to $A^T\rho$ (see, e.g., [1, p. 123] and [10, p. 125]). Thus,

$$\begin{aligned}\bar{U}^T &= \rho A\,\bar{P}^T(L^{-1})^T\\ &= \rho A\,\bar{P}^{(1)}\left(L^{(1)}\right)^T\cdots\bar{P}^{(n-1)}\left(L^{(n-1)}\right)^T\\ &= \rho A\,P^{(1)}U^{(1)}\cdots P^{(n-1)}U^{(n-1)},\end{aligned}$$

by Theorem 4.2 and Remark 4.3. Hence, $A\bar{P}^T(L^{-1})^T = V\rho$, which implies that $A\bar{P}^T = V\rho L^T$, giving the required result with $P = \bar{P}^T$ and $U = L^T$. ∎

We summarize the relationship between GEPP and BDPP in the following corollary.

COROLLARY 4.5. *Suppose A is an n-by-n nonsingular matrix. If the result of applying* GEPP *to A is $\bar{P}A = L\bar{U}$, and the result of applying* BDPP *to ρA^T is $\rho A^T P = V\rho U$, then*

$$\bar{P} = P^T, \quad L = U^T, \quad \textit{and} \quad \bar{U} = \rho V^T\rho.$$

By virtue of the relations between the Bruhat decomposition and the LU decomposition, and between BDPP and GEPP, both Algorithms 2.1 and 4.1 require about $n^3/3$ flops (see, e.g., [1]).

5. Stability of BDPP. For BDPP the growth of entries in U is bounded by 1; thus, from (3.1), the growth factor for BDPP is

$$\gamma_{BP} = \max_{i,j,k}\left|a_{jk}^{(i)}\right| \,/\, \max_{j,k}\left|a_{jk}\right|.$$

For ρW_n, the row reversal of the Wilkinson matrix, it can be shown that $\gamma = 2$, $\gamma_B = 2^{n-1}$, and $\gamma_{BP} = 2$. The transpose of the Wilkinson matrix, W_n^T, is another matrix that has an exponential growth factor ($\gamma_B = 2^{n-1}$) when Algorithm 2.1 is applied and a constant growth factor ($\gamma_{BP} = 4$) when Algorithm 4.1 is applied. Note that by the equivalence in Theorem 4.2, γ for A equals γ_{BP} for ρA^T. Thus, $\gamma_{BP} \le 2^{n-1}$, and this upper bound is realized, for example, by ρW_n^T.

We now show that $\gamma_{BP} \le 2$ for every n-by-n real matrix that has $\gamma = 2^{n-1}$ when GEPP is applied. The following theorem due to Higham and Higham characterizes this class of matrices, which includes W_n.

THEOREM 5.1 (see [6, Theorem 2.2]). *All real n-by-n matrices for which* $\gamma = 2^{n-1}$ *are of the form*

$$A = DM \begin{bmatrix} T & \vdots & \theta d \\ 0 & \vdots & \end{bmatrix},$$

where $D = \mathrm{diag}(\pm 1)$, M *is unit lower triangular with* $m_{ij} = -1$ *for* $i > j$, $T = [t_{ij}]$ *is a nonsingular upper triangular matrix of order* $n-1$, $d = [1 \;\; 2 \;\; 4 \;\; \cdots \;\; 2^{n-1}]^T$, *and* θ *is a scalar such that*

$$\theta = |a_{1n}| = \max_{i,j} |a_{ij}|.$$

For example, the general form of a 5-by-5 matrix with $D = I$ having $\gamma = 2^4$ is

$$A = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & \theta \\ -t_{11} & t_{22} - t_{12} & t_{23} - t_{13} & t_{24} - t_{14} & \theta \\ -t_{11} & -(t_{22} + t_{12}) & t_{33} - (t_{23} + t_{13}) & t_{34} - (t_{24} + t_{14}) & \theta \\ -t_{11} & -(t_{22} + t_{12}) & -(t_{33} + t_{23} + t_{13}) & t_{44} - (t_{34} + t_{24} + t_{14}) & \theta \\ -t_{11} & -(t_{22} + t_{12}) & -(t_{33} + t_{23} + t_{13}) & -(t_{44} + t_{34} + t_{24} + t_{14}) & \theta \end{bmatrix}.$$

THEOREM 5.2. *Let* A *be a real n-by-n matrix for which* $\gamma = 2^{n-1}$ *when* GEPP *is applied. Then application of* BDPP *to* A *gives* $\gamma_{BP} \leq 2$.

Proof. As A is assumed to have $\gamma = 2^{n-1}$, matrix A must be of the form given in Theorem 5.1. At the first step of BDPP on A, if

$$\theta = \max_{1 \leq q \leq n-1} |a_{nq}| = |a_{n1}| = |t_{11}|,$$

then no interchange is performed; however, if

$$\theta > \max_{1 \leq q \leq n-1} |a_{nq}| \quad \text{or} \quad \theta = \max_{2 \leq q \leq n-1} |a_{nq}| = |a_{nk}|$$

with $k \in \{2, \ldots, n-1\}$, then $P^{(1)}$ interchanges columns 1 and n. (Note that this includes a tie-breaking strategy for BDPP.) After one step of Algorithm 4.1,

$$\max_{j,k} \left|a_{jk}^{(1)}\right| \Big/ \max_{j,k} |a_{jk}| \leq 2.$$

The resulting matrix $A^{(1)}$ can be partitioned as

$$A^{(1)} = \begin{bmatrix} z & \vdots & H \\ & \vdots & 0 \end{bmatrix},$$

where z is either column 1 or column n of A, and H is an $(n-1)$-by-$(n-1)$ upper Hessenberg matrix with $h_{i,n-1} = 0$ for $i = 2, \ldots, n-1$. Thus, further steps require only column permutations (but no eliminations). Thus, $\gamma_{BP} \leq 2$. ∎

We conjecture that if an n-by-n nonsingular matrix A can be written as $A = R + xy^T$, where R is an upper triangular matrix, then $\gamma_{BP} \leq 2(n-1)$. The matrices of Theorem 5.1 and the matrices of Foster [3] are of this form.

Acknowledgment. We thank Jaroslaw Oleszczuk for a translation of reference [8] and the referees for constructive comments.

REFERENCES

[1] B. Datta, *Numerical Linear Algebra and Applications*, Brooks/Cole, Toronto, 1995.
[2] L. Elsner, *On some algebraic problems in connection with general eigenvalue algorithms*, Linear Algebra Appl., 48 (1982), pp. 123–138.
[3] L.V. Foster, *Gaussian elimination with partial pivoting can fail in practice*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1354–1362.
[4] I. Gohberg and S. Goldberg, *Finite dimensional Weiner-Hopf equations and factorizations of matrices*, Linear Algebra Appl., 48 (1982), pp. 219–236.
[5] D. Yu. Grigor'ev, *Additive complexity in direct computation*, Theoret. Comput. Sci., 19 (1982), pp. 39–67.
[6] 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.
[7] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[8] L. Yu. Kolotilina, *On the sparsity of factors of the Bruhat decomposition of nonsingular matrices*, Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI), 202 (1992), pp. 5–17 (in Russian). (MR 94m:15008.)
[9] L. Yu. Kolotilina and A. Yu. Yeremin, *Bruhat decomposition and solution of sparse linear algebraic systems*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 421–436.
[10] G.W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[11] J.H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.
[12] S.J. Wright, *A collection of problems for which Gaussian elimination with partial pivoting is unstable*, SIAM J. Sci. Comput., 14 (1993), pp. 231–238.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 99–106, January 1998

007

GENERALIZATIONS OF KY FAN'S DOMINANCE THEOREM*

CHI-KWONG LI† AND ROY MATHIAS†

Abstract. For $x \in \mathbb{R}^n$ and $1 \le k \le n$, define

$$\|x\|_k \equiv \sum_{i=1}^{k} |x_{[i]}|,$$

where $x_{[1]}, \ldots, x_{[n]}$ are the entries of x such that $|x_{[1]}| \ge \cdots \ge |x_{[n]}|$. It is shown that

$$\|x\|^2 \le \|y\|\|z\|$$

for all permutation invariant absolute norms on $\mathbb{R}^n$ if and only if

$$\|x\|_k^2 \le \|y\|_k \|z\|_k, \qquad k = 1, 2, \ldots, n.$$

This generalizes Ky Fan's dominance theorem and implies similar results for unitarily invariant norms on the space of matrices that have application in some recent work of Bhatia, Kittaneh, and Li [*Linear and Multilinear Algebra*, to appear] on inequalities for commutators. Further generalizations of the above result are also obtained.

Key words. permutation invariant absolute norm, singular values

AMS subject classifications. 15A45, 15A60, 15A15

PII. S0895479896311232

1. Introduction and main result.

Given $x \in \mathbb{R}^n$, let $x_{[i]}$ be the ith largest component of x in absolute value. For $k = 1, 2, \ldots, n$, define

$$\|x\|_k \equiv \sum_{i=1}^{k} |x_{[i]}|.$$

Let $\mathbb{R}^n_{+\downarrow}$ denote the set of nonnegative vectors in $\mathbb{R}^n$ with components arranged in nonincreasing order. For any nonzero $\alpha \in \mathbb{R}^n_{+\downarrow}$ define

$$\|x\|_\alpha \equiv \sum_{i=1}^{n} \alpha_i |x_{[i]}|.$$

It is easy to check that both $\|x\|_k$ and $\|x\|_\alpha$ are *permutation invariant absolute norms* (also known as *symmetric gauge functions* or *symmetric norms*; see, e.g., [3, Chapter 3]) on $\mathbb{R}^n$. There is a simple relation between these two families of norms:

$$\|x\|_\alpha = \sum_{k=1}^{n} (\alpha_k - \alpha_{k+1}) \|x\|_k; \tag{1.1}$$

we set $\alpha_{n+1} = 0$. Note that the factors $(\alpha_k - \alpha_{k+1})$ are nonnegative.

Ky Fan's dominance theorem [2, Theorem 7.4.45] is as follows.

* Received by the editors October 25, 1996; accepted for publication (in revised form) by R. Bhatia December 27, 1996.

http://www.siam.org/journals/simax/19-1/31123.html

† Department of Mathematics, The College of William and Mary, Williamsburg, VA 23187 (ckli@math.wm.edu, mathias@math.wm.edu). The research of the second author was partially supported by a grant from the National Science Foundation.

THEOREM 1.1. *Take $x, y \in \mathbb{R}^n$. Then*

$$\|x\| \le \|y\| \tag{1.2}$$

for all permutation invariant absolute norms on $\mathbb{R}^n$ if and only if

$$\|x\|_k \le \|y\|_k, \qquad k = 1, 2, \ldots, n. \tag{1.3}$$

This is a very useful theorem, especially because, according to a result of von Neumann (see, e.g., [2, Theorem 7.4.24]), any *unitarily invariant norm* on the space of matrices can be represented as a permutation invariant absolute norm of the singular values of the matrix. For $1 \le k \le n$, define the *Ky Fan k-norm* of an $n \times n$ matrix X by

$$\|X\|_k = \sum_{i=1}^{k} \sigma_i(X),$$

where $\sigma_1(X) \ge \cdots \ge \sigma_n(X)$ denote the singular values of X. Then Theorem 1.1 immediately implies a similar result for unitarily invariant norms.

THEOREM 1.2. *Take $n \times n$ complex matrices X and Y. Then*

$$\|X\| \le \|Y\|$$

for all unitarily invariant norms on the space of $n \times n$ complex matrices if and only if

$$\|X\|_k \le \|Y\|_k, \qquad k = 1, 2, \ldots, n.$$

In this note we give some generalizations of Theorem 1.1. These generalizations imply generalizations of Theorem 1.2 for unitarily invariant norms. One application of Theorem 1.4 is in the work of Bhatia, Kittaneh, and Li [1] on inequalities for commutators: Theorem 1.4 here implies that the bound [1, equation (2.3)] for the Ky Fan k-norms is valid for all unitarily invariant norms.

To prove Theorems 1.4 and 2.1 we use the following quasi-linear representation of permutation invariant absolute norms that is essentially [4, Theorem 2.1].

THEOREM 1.3. *Let $\|\cdot\|$ be a permutation invariant absolute norm on $\mathbb{R}^n$. Then there is a compact convex set $\mathcal{A} \subseteq \mathbb{R}^n_{+\downarrow}$ such that for all $x \in \mathbb{R}^n$*

$$\|x\| = \max\{\|x\|_\alpha : \alpha \in \mathcal{A}\}.$$

In some sense, the above result says that the collection of $\|\cdot\|_\alpha$ with $\alpha \in \mathbb{R}^n_{+\downarrow}$ forms a generating set for permutation invariant absolute norms. In fact, the set $\mathcal{A}$ in Theorem 1.3 can be taken to be the intersection of $\mathbb{R}^n_{+\downarrow}$ and the unit ball of the dual of $\|\cdot\|$. To illustrate the power of this representation we use it to prove Theorem 1.1.

Let $x, y \in \mathbb{R}^n$ satisfy (1.3). Choose a permutation invariant absolute norm and let $\mathcal{A}$ be a corresponding set. Let $\alpha \in \mathcal{A}$ be such that

$$\|x\| = \|x\|_\alpha.$$

Then

$$\|x\| = \|x\|_\alpha$$

$$\begin{aligned}
&= \sum_{k=1}^{n} (\alpha_k - \alpha_{k+1}) \|x\|_k \\
&\leq \sum_{k=1}^{n} (\alpha_k - \alpha_{k+1}) \|y\|_k \\
&= \|y\|_\alpha \\
&\leq \|y\|.
\end{aligned}$$

This proof is considerably simpler than the standard proof which involves doubly stochastic matrices and Birkhoff's theorem; see, e.g., [2, proof of Theorem 7.4.45]. The proof that we have just given is essentially the same as [3, proof of Corollary 3.5.9].

The following is our main result. Although we shall further generalize it in the next section, we prefer to present the statement and proof here since the statement can be directly applied to the work of Bhatia, Kittaneh, and Li [1, Theorem 2.2] on spectral variation as mentioned before, and the proof contains one of the key ideas in this paper. To see that Theorem 1.4 is indeed a generalization of Theorem 1.1, take $y = z$ and take square roots of (1.4) and (1.5).

THEOREM 1.4. *Let $x, y, z \in \mathbb{R}^n$. Then*

$$\|x\|^2 \leq \|y\|\|z\| \tag{1.4}$$

for all permutation invariant absolute norms on $\mathbb{R}^n$ if and only if

$$\|x\|_k^2 \leq \|y\|_k \|z\|_k, \qquad k = 1, 2, \ldots, n. \tag{1.5}$$

Consequently, for any $n \times n$ matrices X, Y, Z,

$$\|X\|^2 \leq \|Y\|\|Z\|$$

for all unitarily invariant norms if and only if

$$\|X\|_k^2 \leq \|Y\|_k \|Z\|_k, \qquad k = 1, 2, \ldots, n.$$

Proof. Clearly, one only needs to prove the ($\Leftarrow$) part. Take any permutation invariant absolute norm $\|\cdot\|$ on $\mathbb{R}^n$. Let $\mathcal{A}$ be a compact convex set corresponding to the norm $\|\cdot\|$. Let $\alpha \in \mathcal{A}$ be such that

$$\|x\|_\alpha = \|x\|.$$

The condition (1.5) ensures that the 2×2 matrices

$$\begin{pmatrix} \|y\|_k & \|x\|_k \\ \|x\|_k & \|z\|_k \end{pmatrix}, \qquad k = 1, 2, \ldots, n,$$

are positive semidefinite. Since the quantities $(\alpha_k - \alpha_{k+1})$ are nonnegative it follows that

$$\sum_{k=1}^{n} (\alpha_k - \alpha_{k+1}) \begin{pmatrix} \|y\|_k & \|x\|_k \\ \|x\|_k & \|z\|_k \end{pmatrix} = \begin{pmatrix} \|y\|_\alpha & \|x\|_\alpha \\ \|x\|_\alpha & \|z\|_\alpha \end{pmatrix}$$

is also positive semidefinite. Using the nonnegativity of the determinant of this matrix for the first inequality and the representation of $\|\cdot\|$ in terms of $\mathcal{A}$ for the second we have

$$\|x\|^2 = \|x\|_\alpha^2 \leq \|y\|_\alpha \|z\|_\alpha \leq \|y\|\|z\|$$

which is the required inequality. ∎

2. Generalizations. One can extend Theorem 1.4 to several vectors and prove that

(i) *for $x_0, x_1, \ldots, x_m \in \mathbb{R}^n$, we have*

$$\|x_0\|^m \le \prod_{i=1}^{m} \|x_i\| \tag{2.1}$$

for all permutation invariant absolute norms on $\mathbb{R}^n$ if and only if

$$\|x_0\|_k^m \le \prod_{i=1}^{m} \|x_i\|_k, \qquad k = 1, 2, \ldots, n. \tag{2.2}$$

A similar (and actually simpler) proof using the quasi-linear representation in Theorem 1.3 yields

(ii) *for $x_0, x_1, \ldots, x_m \in \mathbb{R}^n$, we have*

$$\|x_0\| \le \frac{1}{m} \sum_{i=1}^{m} \|x_i\| \tag{2.3}$$

for all permutation invariant absolute norms on $\mathbb{R}^n$ if and only if

$$\|x_0\|_k \le \frac{1}{m} \sum_{i=1}^{m} \|x_i\|_k, \qquad k = 1, 2, \ldots, n. \tag{2.4}$$

(iii) *For nonzero $x_0, x_1, \ldots, x_m \in \mathbb{R}^n$, we have*

$$\|x_0\| \le m \left\{ \sum_{i=1}^{m} \|x_i\|^{-1} \right\}^{-1} \tag{2.5}$$

for all permutation invariant absolute norms on $\mathbb{R}^n$ if and only if

$$\|x_0\|_k \le m \left\{ \sum_{i=1}^{m} \|x_i\|_k^{-1} \right\}^{-1}, \qquad k = 1, 2, \ldots, n. \tag{2.6}$$

More generally, we have the following theorem.

THEOREM 2.1. *Suppose $f : \mathbb{R}_+^m \to \mathbb{R}$ is concave, increasing in each variable and homogeneous, i.e., $f(ta_1, \ldots, ta_m) = tf(a_1, \ldots, a_m)$ for any $t \ge 0$. Then*

$$\|x_0\| \le f(\|x_1\|, \ldots, \|x_m\|) \tag{2.7}$$

for all permutation invariant absolute norms $\|\cdot\|$ on $\mathbb{R}^n$ if and only if

$$\|x_0\|_k \le f(\|x_1\|_k, \ldots, \|x_m\|_k), \qquad k = 1, 2, \ldots, n. \tag{2.8}$$

Consequently, for any $n \times n$ matrices $X_0, X_1, \ldots, X_m$,

$$\|X_0\| \le f(\|X_1\|, \ldots, \|X_m\|)$$

for all unitarily invariant norms $\|\cdot\|$ if and only if

$$\|X_0\|_k \le f(\|X_1\|_k, \ldots, \|X_m\|_k), \qquad k = 1, 2, \ldots, n.$$

Proof. We need to prove only the ($\Leftarrow$) part. Take any permutation invariant absolute norm $\|\cdot\|$ on $\mathbb{R}^n$. Let $\mathcal{A}$ be a compact convex set corresponding to the norm $\|\cdot\|$. Then there exists $\alpha \in \mathcal{A}$ such that $\|x_0\| = \|x_0\|_\alpha$. Let $\beta_k = (\alpha_k - \alpha_{k+1})/\alpha_1$. Then

$$\begin{aligned}
\|x_0\| &= \|x_0\|_\alpha \\
&= \sum_{k=1}^{n} \beta_k (\alpha_1 \|x_0\|_k) \\
&\le \sum_{k=1}^{n} \beta_k f(\alpha_1 \|x_1\|_k, \ldots, \alpha_1 \|x_n\|_k) \\
&\le f\left(\sum_{k=1}^{n} \beta_k \alpha_1 \|x_1\|_k, \ldots, \sum_{k=1}^{n} \beta_k \alpha_1 \|x_n\|_k \right) \\
&= f(\|x_1\|_\alpha, \ldots, \|x_n\|_\alpha) \\
&\le f(\|x_1\|, \ldots, \|x_n\|).
\end{aligned}$$

We have used the homogeneity of f and (2.8) for the first inequality and the concavity of f for the second, and the increasing property of f for the final inequality. ∎

To verify the concavity of the geometric and harmonic means one can compute the Hessian and show that it is negative semidefinite on $\mathbb{R}^n_+$.

Note that Theorem 2.1 is very similar to [3, Corollary 3.5.11], which asserts that (2.7) holds for all functions f that increase in each variable if and only if (2.7) is valid for all α-norms $\|\cdot\|_\alpha$. Our result focuses on those functions f satisfying (2.7) whenever the finite set of conditions in (2.8) hold, and is easier to use in applications (cf. [1, Theorem 2.2]). There are other types of norms that admit quasi-linear representations (e.g., see [5, Theorem 3.3]) so that one may prove results similar to [3, Corollary 3.5.11] for such norms. However, in many cases, it is impossible to obtain results similar to that of Ky Fan (e.g., see [5, section 4]), and hence hopeless to obtain analogues of Theorem 2.1.

Another direction to extend the result of Ky Fan is to consider the functions $\phi : \mathbb{R}^n_+ \to \mathbb{R}$ such that

$$\phi(x) \le \phi(y) \quad \text{whenever (1.3) holds.}$$

This has been done extensively in connection with the theory of *majorization*; e.g., see [6]. In view of our Theorem 1.4 and statement (i), it is natural to consider the set $\mathcal{P}_m$ of functions $\phi : \mathbb{R}^n_+ \to \mathbb{R}$ such that

$$\phi(x_0)^m \le \prod_{i=1}^{m} \phi(x_i) \quad \text{whenever (2.2) holds.}$$

By the previous results, it is clear that for each $m = 1, 2, \ldots$, the set $\mathcal{P}_m$ contains all permutation invariant absolute norms. Also, it is not hard to show that

$$\mathcal{P}_1 \supseteq \mathcal{P}_2 \supseteq \mathcal{P}_3 \supseteq \cdots. \tag{2.9}$$

Evidently (e.g., see [6, Chapter 3]), a function $\phi : \mathbb{R}^n_+ \to \mathbb{R}$ belongs to $\mathcal{P}_1$ if and only if the function $\tilde{\phi}$ that satisfies

$$\phi(x) = \tilde{\phi}(\|x\|_n, \ldots, \|x\|_1)$$

is increasing in each variable on the domain

$$\mathcal{D} = \{\tilde{x} = (\|x\|_n, \ldots, \|x\|_1) : x \in \mathbb{R}_+^n\}.$$

In the same spirit, one sees that a function $\phi : \mathbb{R}_+^n \to \mathbb{R}$ belongs to $\mathcal{P}_2$ if and only if the function $\tilde{\phi}$ defined above satisfies

$$\tilde{\phi}(x)^2 \le \tilde{\phi}(y)\tilde{\phi}(z) \quad \text{whenever } x, y, z \in \mathcal{D} \text{ satisfy } y \circ z - x \circ x \in \mathbb{R}_+^n, \tag{2.10}$$

where $\circ$ denotes the Schur (entrywise) product of vectors. This condition is reasonably easy to check, and it is not difficult to construct examples of $\tilde{\phi}$ that do not correspond to permutation invariant absolute norms. For instance, one may let $\tilde{\phi}$ be the kth elementary symmetric function or the kth completely symmetric function on n variables for any $1 \le k \le n$. Furthermore, one may construct $\tilde{\phi}$ which is increasing in each variable, but (2.10) does not hold. One will then get a function $\phi \in \mathcal{P}_1 \setminus \mathcal{P}_2$. For example, if $\phi(x) = \log(1 + \|x\|_1)$, then clearly $\phi \in \mathcal{P}_1$, but (2.10) does not hold for $x = (2, 0, \ldots, 0)$, $y = 2x$, and $z = x/2$.

The characterization of functions in $\mathcal{P}_m$ becomes more complicated and not so easy to check if $m \ge 3$. As a result, it is difficult to use the technique in the preceding paragraph to check whether the inclusion

$$\mathcal{P}_m \supseteq \mathcal{P}_{m+1}$$

is proper for $m \ge 2$. Fortunately, we have the following result.

THEOREM 2.2. *For each positive integer m, let $\mathcal{P}_m$ be the collection of $\phi : \mathbb{R}_+^n \to \mathbb{R}$ such that*

$$\phi(x_0)^m \le \prod_{i=1}^m \phi(x_i) \quad \textit{whenever (2.2) holds.} \tag{2.11}$$

Then

$$\mathcal{P}_1 \supsetneq \mathcal{P}_2 = \mathcal{P}_3 = \mathcal{P}_4 = \cdots.$$

Proof. By the previous discussion and (2.9), it suffices to prove that $\mathcal{P}_2 \subseteq \mathcal{P}_m$. Suppose $\phi \in \mathcal{P}_2$, and $x_0, x_1, \ldots, x_m \in \mathbb{R}_+^n$. We may assume that $x_i \in \mathbb{R}_{+\downarrow}^n$ for all $i = 0, \ldots, m$. If $x_1 = \cdots = x_m$, then $\|x_0\|_k \le \|x_1\|_k$ for all $k = 1, \ldots, m$. Since $\phi \in \mathcal{P}_2 \subseteq \mathcal{P}_1$, it follows that $\phi(x_0)^m \le \phi(x_1)^m = \prod_{i=1}^m \phi(x_i)$.

Suppose not all x_i are equal for $i = 1, \ldots, m$. Pick indices p and q such that

$$\|x_p - x_q\|_n = \max\{\|x_r - x_s\|_n : 1 \le r < s \le m\}.$$

Let $u_0 = v_0 = w_0 = 0$, $u_k = \|x_p\|_k$, $v_k = \|x_q\|_k$, $w_k = \sqrt{u_k v_k}$ for $k = 1, \ldots, n$. Then $u_{r+1} - u_r \le u_r - u_{r-1}$ and $v_{r+1} - v_r \le v_r - v_{r-1}$ for $r = 1, \ldots, n-1$. It follows that

$$\begin{aligned} w_{r+1} + w_{r-1} &= \sqrt{u_{r+1}v_{r+1}} + \sqrt{u_{r-1}v_{r-1}} \\ &\le \{(u_{r+1} + u_{r-1})(v_{r+1} + v_{r-1})\}^{1/2} \\ &\le 2w_r, \end{aligned}$$

i.e., $w_{r+1} - w_r \le w_r - w_{r-1}$, for $r = 1, \ldots, n-1$. Define

$$\tilde{x}_p = \tilde{x}_q = (w_1 - w_0, w_2 - w_1, \ldots, w_n - w_{n-1}) \in \mathbb{R}_{+\downarrow}^n.$$

We have

$$(\|\tilde{x}_p\|_n, \ldots, \|\tilde{x}_p\|_1) \circ (\|\tilde{x}_q\|_n, \ldots, \|\tilde{x}_q\|_1) = (\|x_p\|_n, \ldots, \|x_p\|_1) \circ (\|x_q\|_n, \ldots, \|x_q\|_1).$$

Since $\phi \in \mathcal{P}_2$, we have $\phi(\tilde{x}_p)^2 \le \phi(x_p)\phi(x_q)$, and hence

$$\prod_{i=1}^{m} \phi(\tilde{x}_i) \le \prod_{i=1}^{m} \phi(x_i)$$

if $\tilde{x}_r = x_r$ for all $r \ne p, q$. Moreover, the equality holds if ϕ is replaced by $\|\cdot\|_k$ for each $k = 1, \ldots, n$.

Iterating the above procedure, we see that the m vectors will converge to a single vector, say $\tilde{x}$, and this $\tilde{x}$ will satisfy

$$\phi(\tilde{x})^m \le \prod_{i=1}^{m} \phi(x_i),$$

and the equality will hold if ϕ is replaced by $\|\cdot\|_k$ for $k = 1, \ldots, n$. Thus

$$\|x_0\|_k^m \le \prod_{i=1}^{m} \|x_i\|_k = \|\tilde{x}\|_k^m, \qquad k = 1, \ldots, n.$$

Since $\phi \in \mathcal{P}_2 \subseteq \mathcal{P}_1$, we have

$$\phi(x_0) \le \phi(\tilde{x})$$

and hence

$$\phi(x_0)^m \le \phi(\tilde{x})^m \le \prod_{i=1}^{m} \phi(x_i)$$

as required. □

Similarly, one may consider $\mathcal{S}_m$ to be the set of functions $\phi : \mathbb{R}_+^n \to \mathbb{R}$ such that

$$\phi(x_0) \le \frac{1}{m} \sum_{i=1}^{m} \phi(x_i) \quad \text{whenever (2.4) holds,}$$

and consider $\mathcal{H}_m$ to be the set of functions $\phi : \mathbb{R}_+^n \to \mathbb{R}$ such that

$$\phi(x_0) \le m \left\{ \sum_{i=1}^{m} \phi(x_i)^{-1} \right\}^{-1} \quad \text{whenever (2.6) holds.}$$

Again, one can show that

$$\mathcal{S}_1 \supsetneq \mathcal{S}_2 = \mathcal{S}_3 = \mathcal{S}_4 = \cdots \quad \text{and} \quad \mathcal{H}_1 \supsetneq \mathcal{H}_2 = \mathcal{H}_3 = \mathcal{H}_4 = \cdots.$$

In fact, if $\phi(x) = \|x\|_1^{1/2}$, then $\phi \in \mathcal{S}_1$, but $\phi \notin \mathcal{S}_2$ as $\phi(x) > (\phi(y) + \phi(z))/2$ with $x = 3v, y = 2v$, and $z = 4v$ for any nonzero $v \in \mathbb{R}^n$. Similarly, if $\phi(x) = \|x\|_1^2$, then $\phi \in \mathcal{H}_1$, but $\phi \notin \mathcal{H}_2$ as $\phi(x) > 2\{(\phi(y)^{-1} + \phi(z)^{-1}\}^{-1}$ with $x = v/3, y = v/2$, and $z = v/4$ for any nonzero $v \in \mathbb{R}^n$.

One may also consider generalizations along the direction of Theorem 2.1.

REFERENCES

[1] R. Bhatia, F. Kittaneh, and R.-C. Li, *Some inequalities for commutators and an application to spectral variation* II, Linear and Multilinear Algebra, to appear.
[2] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.
[3] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.
[4] R. A. Horn and R. Mathias, *Cauchy-Schwarz inequalities associated with positive semidefinite matrices*, Linear Algebra Appl., 142 (1990), pp. 63–82.
[5] C.-K. Li and N.-K. Tsing, *G-invariant norms and G(c)-radii*, Linear Algebra Appl., 150 (1991), pp. 179–194.
[6] A. W. Marshall and I. Olkin, *Inequalities: The Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 107–139, January 1998

008

A FAST STABLE SOLVER FOR NONSYMMETRIC TOEPLITZ AND QUASI-TOEPLITZ SYSTEMS OF LINEAR EQUATIONS*

S. CHANDRASEKARAN† AND ALI H. SAYED‡

Abstract. We derive a stable and fast solver for nonsymmetric linear systems of equations with shift structured coefficient matrices (e.g., Toeplitz, quasi-Toeplitz, and product of two Toeplitz matrices). The algorithm is based on a modified fast QR factorization of the coefficient matrix and relies on a stabilized version of the generalized Schur algorithm for matrices with displacement structure. All computations can be done in $O(n^2)$ operations, where n is the matrix dimension, and the algorithm is backward stable.

Key words. displacement structure, generalized Schur algorithm, QR factorization, hyperbolic rotations, generator matrices, Schur complements, error analysis

AMS subject classifications. 65F05, 65G05, 65F30, 15A23

PII. S0895479895296458

1. Introduction.

Linear systems of equations can be solved by resorting to the LDU factorization (Gaussian elimination) of the coefficient matrix. But for indefinite or nonsymmetric matrices, the LDU factorization is numerically unstable if done without pivoting. Moreover, since pivoting can destroy the structure of a matrix, it is not always possible to incorporate it into a fast algorithm for structured matrices without potential loss of computational efficiency.

Sometimes though, one can transform a given structured matrix to another structured form so that the new structure is insensitive to *partial* pivoting operations [9, 12]. While this technique can be satisfactory for certain situations, it may still pose numerical problems because partial pivoting by itself is not sufficient to guarantee numerical stability even for slow algorithms. It also seems difficult to implement *complete* pivoting in a fast algorithm without accruing a considerable loss of efficiency. Recently, Gu [11] proposed a fast algorithm that incorporates an *approximate* complete pivoting strategy.

Another way to solve a structured linear system of equations is to compute the QR factorization of the coefficient matrix rapidly. Several fast methods have been proposed earlier in the literature [1, 6, 7, 8, 19], but none of them are numerically stable.

In this paper we resolve this open issue and derive an algorithm that is provably both fast *and* backward stable for solving linear systems of equations involving nonsymmetric structured coefficient matrices (e.g., Toeplitz, quasi Toeplitz, and Toeplitz-like). The algorithm is based on a modified fast QR factorization of the coefficient matrix T in $Tx = b$. It computes a factorization for T of the form

$$T = \Delta(\Delta^{-1}Q)R,$$

*Received by the editors December 27, 1995; accepted for publication (in revised form) by L. Reichel January 4, 1997.
http://www.siam.org/journals/simax/19-1/29645.html
†Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 (shiv@ece.ucsb.edu).
‡Department of Electrical Engineering, University of California, Los Angeles, CA 90095 (sayed@ee.ucla.edu). The work of this author was supported in part by National Science Foundation award MIP-9796147.

where Δ is lower triangular, $(\Delta^{-1}Q)$ is orthogonal, and R is upper triangular. The factorization is then used to solve for x efficiently by using

$$x = R^{-1}(Q^T\Delta^{-T})\Delta^{-1}b. \tag{1.1}$$

All computations can be done in $O(n^2)$ operations, where n is the matrix dimension, and the algorithm is backward stable in the sense that the computed solution $\hat{x}$ is shown to satisfy an equation of the form

$$(T + H)\hat{x} = b,$$

where the norm of the error matrix satisfies

$$\|H\| \ \leq \ c_1\epsilon \, \|T\| \ + \ O(\epsilon^2),$$

where ϵ denotes machine precision and c_1 is a low-order polynomial in n.

The fast and stable algorithm to be derived in this paper is based on ideas of displacement structure theory [15]. The concept of displacement structure was introduced by Kailath, Kung, and Morf almost two decades ago [14] and has since proven to be a useful tool in matrix analysis. Its strength lies in the fact that it allows us, in a systematic way, to describe and exploit varied forms of matrix structure. In this framework, matrix structures are described in terms of *displacement equations* and triangular factorizations are efficiently carried out by a *generalized Schur algorithm* [15].

However, the numerical behavior of the generalized Schur algorithm has been an issue of concern until very recently, which is mainly due to the fact that the algorithm relies heavily on hyperbolic transformations. In recent work, Bojanczyk et al. [2] have shown that for a subclass of positive-definite shift structured matrices (known as quasi Toeplitz), the Cholesky factorization provided by the generalized Schur algorithm is asymptotically stable despite the hyperbolic rotations.

The class of quasi-Toeplitz matrices refers to a special kind of structured matrices whose displacement rank (to be defined later) is equal to 2. Stewart and van Dooren [18] further considered the case of positive-definite shift structured matrices with displacement ranks larger than 2. They argued that the generalized Schur algorithm will still provide a stable Cholesky factorization provided the required rotations are now implemented in a special way (a combination of unitary rotations followed by a single hyperbolic rotation in mixed form).

Motivated by the work of Bojanczyk et al. [2], we have also pursued in [4] a detailed analysis of the numerical stability of the generalized Schur algorithm for a general class of positive-definite structured matrices. In particular, we have shown that along with proper implementations of the hyperbolic transformations, if further modifications are introduced while computing intermediate quantities, the algorithm will guarantee a Cholesky factorization that is provably backward stable. We further employed a perturbation analysis to indicate the best accuracy that can be expected from any finite precision algorithm (slow or fast), and then showed that the modified Schur algorithm of [4] essentially achieves this bound. For all practical purposes, the major conclusion of the analysis in [4] was that the modified Schur algorithm is backward stable for a large class of structured matrices.

The above results have further motivated us to tackle the standing issue of deriving an algorithm that is both *fast and stable* for the solution of *nonsymmetric* structured linear systems of equations $Tx = b$, where T is shift structured (to be defined later). The stability analyses of the generalized Schur algorithm that we referred

to above do not apply in this case since the structured matrix T is not positive definite (it is not even required to be symmetric). The only restriction on T is invertibility.

The way we approach the problem is motivated by embedding ideas pursued in [5, 13]. We first embed the given $n \times n$ matrix T into a larger $2n \times 2n$ matrix M that is defined by

$$M = \begin{bmatrix} T^T T & T^T \\ T & \mathbf{0} \end{bmatrix}. \tag{1.2}$$

The matrix M is symmetric but still indefinite; while its leading $n \times n$ submatrix is positive definite (equal to $T^T T$), its Schur complement with respect to the $(1,1)$ block is negative definite (and equal to $-I$). (The product $T^T T$ is not formed explicitly, as explained later.)

We then apply $2n$ steps of the generalized Schur algorithm to M and obtain its *computed* triangular factorization, which is of the form

$$\begin{bmatrix} \hat{R}^T & \mathbf{0} \\ \hat{Q} & \Delta \end{bmatrix} \begin{bmatrix} \hat{R} & \hat{Q}^T \\ \mathbf{0} & -\Delta^T \end{bmatrix},$$

where $\hat{R}^T$ and Δ are $n \times n$ lower triangular matrices. The matrices $\{\hat{R}, \hat{Q}, \Delta\}$ are the quantities used in (1.1) to determine the computed solution $\hat{x}$ in a backward stable manner.

From a numerical point of view, the above steps differ in crucial ways from the embeddings suggested in [5, 13], and which turn out to mark the difference between a numerically stable and a numerically unstable implementation.

The discussion in [5, pp. 37, 50, 52] and [13] is mainly concerned with fast procedures for the QR factorization of Toeplitz-block and block-Toeplitz matrices. It employs an embedding of the form

$$M = \begin{bmatrix} T^T T & T^T \\ T & I \end{bmatrix}, \tag{1.3}$$

where the identity matrix I in (1.3) replaces the zero matrix in our embedding (1.2). The derivation in [5, 13] suggests applying n (rather than $2n$) steps of the generalized Schur algorithm to (1.3) and then uses the resulting $\hat{R}$ and $\hat{Q}$ as the QR factors of T. This procedure, however, *does not* guarantee a numerically orthogonal matrix $\hat{Q}$ and cannot, therefore, be used to implement a stable solver for a linear system of equations $Tx = b$.

For this reason, we instead propose in this paper to proceed with the earlier embedding (1.2) since it seems difficult to obtain a stable algorithm that is solely based on the alternative embedding (1.3). We also apply $2n$ steps (rather than just n steps) of the generalized Schur algorithm to (1.2). This allows us to incorporate a correction procedure into the algorithm that is shown to ensure backward stability, when coupled with other modifications that are needed, especially while applying the hyperbolic rotations.

1.1. Notation. In the discussion that follows we use $\|\cdot\|$ to denote the 2-norm of its argument. Also, the $\hat{}$ notation denotes computed quantities, and we use ϵ to 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, multiplications, 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 $|\delta| \leq \epsilon$. Likewise 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 matrix M and an $n \times n$ lower triangular real-valued matrix F. The displacement of M with respect to F is denoted by ∇_F and defined as

$$\nabla_F = M - FMF^T. \tag{2.1}$$

The matrix M is said to have low displacement rank with respect to F if the rank of ∇_F is considerably lower than n. In this case, M is said to have displacement structure with respect to F [15].

Let $r \ll n$ denote the rank of ∇_F. It follows that we can factor ∇_F as

$$\nabla_F = GJG^T, \tag{2.2}$$

where G is an $n \times r$ matrix and J is a signature matrix of the form

$$J = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}, \quad p + q = r. \tag{2.3}$$

The integer p denotes the number of positive eigenvalues of ∇_F, while the integer q denotes the number of its negative eigenvalues. The factorization (2.2) is highly nonunique. If G satisfies (2.2), then $G\Theta$ also satisfies (2.2) 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 (2.1) and (2.2), a matrix M is said to be structured with respect to the displacement operation defined by (2.1) if it satisfies a displacement equation of the form

$$M - FMF^T = GJG^T, \tag{2.4}$$

with a "low" rank matrix G. Equation (2.4) uniquely defines M (i.e., it has a unique solution M) iff 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 hold for the cases studied in this paper. (It can be relaxed in some instances [15].)

The pair (G, J) is said to be a generator pair for M since, along with F, it completely identifies M. Note, however, that while M 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 M, will generally be an order of magnitude faster than algorithms that operate on the entries of M 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 triangular factorization of M. A recent survey on various other forms of displacement structure and on the associated forms of Schur algorithms can be found in [15].

The notion of structured matrices can also be extended to nonsymmetric matrices M. In this case, the displacement of M is generally defined with respect to two lower triangular matrices F and A (which can be the same, i.e., $F = A$; see (2.10)),

$$\nabla_{F,A} = M - FMA^T, \tag{2.5}$$

and the low-rank difference matrix $\nabla_{F,A}$ is (nonuniquely) factored as

$$\nabla_{F,A} = GB^T, \tag{2.6}$$

where G and B are $n \times r$ generator matrices, i.e.,

$$M - FMA^T = GB^T. \tag{2.7}$$

Again, this displacement equation uniquely defines M iff the diagonal entries of F and A satisfy $1 - f_i a_j \neq 0$ for all i, j, a condition that will be met in this paper.

2.1. Toeplitz, quasi-Toeplitz, and shift structured matrices. The concept of displacement structure is perhaps best introduced by considering the much-studied special case of a symmetric Toeplitz matrix $T = \left[t_{|i-j|}\right]_{i,j=1}^{n}$, $t_0 = 1$.

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):

$$Z = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}. \tag{2.8}$$

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

$$T - ZTZ^T = \begin{bmatrix} 1 & 0 \\ t_1 & t_1 \\ \vdots & \vdots \\ t_{n-1} & t_{n-1} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ t_1 & t_1 \\ \vdots & \vdots \\ t_{n-1} & t_{n-1} \end{bmatrix}^T = GJG^T. \tag{2.9}$$

Similarly, for a nonsymmetric Toeplitz matrix $T = [t_{i-j}]_{i,j=1}^{n}$, we can easily verify that the difference $T - ZTZ^T$ has displacement rank 2 and that a generator (G, B) for T is

$$T - ZTZ^T = \begin{bmatrix} t_0 & 1 \\ t_1 & 0 \\ \vdots & \vdots \\ t_{n-1} & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & t_{-1} \\ \vdots & \vdots \\ 0 & t_{-n+1} \end{bmatrix}^T = GB^T. \tag{2.10}$$

This is a special case of (2.7) with $F = A = Z$. In particular, any matrix T for which $(T - ZTZ^T)$ has rank 2 is called *quasi Toeplitz*, i.e.,

$$T - ZTZ^T = GB^T \text{ has rank 2.} \tag{2.11}$$

For example, the inverse of a Toeplitz matrix is quasi Toeplitz [15].

Later in the paper we shall focus on the class of *shift structured* matrices (cf. (4.1)), which includes Toeplitz and quasi-Toeplitz matrices as special cases. These are all matrices that are structured with respect to $F = A = Z$. For ease of reference, we define the terminology below.

DEFINITION 2.1.

1. *Any matrix that is structured with respect to the shift operators $F = Z$ and $A = Z$ will be said to be shift structured. That is, for shift structured matrices the rank of $\nabla_{Z,Z}$ (or displacement rank) is low compared to n.*
2. *A quasi-Toeplitz matrix is a shift structured matrix with displacement rank* 2.

For example, the product of two Toeplitz matrices is shift structured with displacement rank 4 [15].

3. The generalized Schur algorithm. An efficient algorithm for the triangular factorization of symmetric or nonsymmetric structured matrices (of either forms (2.4) or (2.7)) is the generalized Schur algorithm [15]. For our purposes, it is sufficient to describe the algorithm here for *symmetric* structured matrices M of the form (2.4), with a *strictly* lower triangular matrix F. This includes, for example, the following special choices for F: $F = Z$, $F = Z^2$, $F = (Z \oplus Z)$, etc. The matrix M is further assumed to be strongly regular (i.e., all its leading submatrices are nonsingular).

A generator matrix G is said to be in *proper* form if its first nonzero row has a single nonzero entry, say in the first column

$$G = \begin{bmatrix} x & 0 & 0 & 0 & 0 \\ x & x & x & x & x \\ x & x & x & x & x \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x & x & x & x & x \end{bmatrix}, \tag{3.1}$$

or in the last column

$$G = \begin{bmatrix} 0 & 0 & 0 & 0 & x \\ x & x & x & x & x \\ x & x & x & x & x \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x & x & x & x & x \end{bmatrix}. \tag{3.2}$$

The generalized Schur algorithm operates on the entries of (F, G, J), which describe the displacement structure of M in (2.4) (assumed strongly regular), and provides the triangular factorization of M [15].

ALGORITHM 3.1 (the generalized Schur algorithm).

- *Input data: An $n \times n$ strictly lower triangular matrix F, an $n \times r$ generator $G_1 = G$, and $J = (I_p \oplus -I_q)$.*
- *Output data: A lower triangular factor L and a signature matrix D such that $M = LDL^T$, where M is the solution of* (2.4) *(assumed $n \times n$).*

The algorithm operates as follows: start with $G_1 = G$, $F_1 = F$, *and repeat for* $i = 1, 2, \ldots, n$:

1. *Let* g_i *denote the top row of* G_i.
2. *If* $g_i J g_i^T > 0$ *(we refer to this case as a positive step):*
   - *Choose a* J*-unitary rotation* Θ_i *that converts* g_i *to proper form with respect to the first column, i.e.,*

$$g_i \Theta_i = \begin{bmatrix} x & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{3.3}$$

   Let $\bar{G}_i = G_i \Theta_i$ *(i.e., apply* Θ_i *to* G_i*).*
   - *The nonzero part of the* i*th column of* L*, denoted by* $\bar{l}_i$*, is the first column of* $\bar{G}_i$*,*

$$\bar{l}_i = \bar{G}_i \begin{bmatrix} 1 \\ \mathbf{0} \end{bmatrix}. \tag{3.4}$$

   The i*th column of* L*, denoted by* l_i*, is obtained by appending* $(i-1)$ *zero entries to* $\bar{l}_i$*,*

$$l_i = \begin{bmatrix} \mathbf{0} \\ \bar{l}_i \end{bmatrix}. \tag{3.5}$$

   The i*th signature is* $d_i = 1$.
   - *Keep the last columns of* $\bar{G}_i$ *unchanged and multiply the first column by* F_i*, where* F_i *denotes the submatrix obtained by deleting the first* $(i-1)$ *rows and columns of* F*. This provides a new matrix whose first row is zero (since* F_i *is strictly lower triangular) and whose last rows are the rows of the next generator matrix* G_{i+1}*, i.e.,*

$$\begin{bmatrix} \mathbf{0} \\ G_{i+1} \end{bmatrix} = \begin{bmatrix} F_i \bar{l}_i & \bar{G}_i \begin{bmatrix} 0 & \\ & I \end{bmatrix} \end{bmatrix}. \tag{3.6}$$

3. *If* $g_i J g_i^T < 0$ *(we refer to this case as a negative step):*
   - *Choose a* J*-unitary rotation* Θ_i *that converts* g_i *to proper form with respect to the last column, i.e.,*

$$g_i \Theta_i = \begin{bmatrix} 0 & 0 & 0 & 0 & x \end{bmatrix}. \tag{3.7}$$

   Let $\bar{G}_i = G_i \Theta_i$ *(i.e., apply* Θ_i *to* G_i*).*
   - *The nonzero part of the* i*th column of* L*, denoted by* $\bar{l}_i$*, is the last column of* $\bar{G}_i$*,*

$$\bar{l}_i = \bar{G}_i \begin{bmatrix} \mathbf{0} \\ 1 \end{bmatrix}. \tag{3.8}$$

   The i*th column of* L*, denoted by* l_i*, is obtained by appending* $(i-1)$ *zero entries to* $\bar{l}_i$*,*

$$l_i = \begin{bmatrix} \mathbf{0} \\ \bar{l}_i \end{bmatrix}. \tag{3.9}$$

   The i*th signature is* $d_i = -1$.

- *Keep the first columns of $\bar{G}_i$ unchanged and multiply the last column by F_i. This provides a new matrix whose first row is zero (since F_i is strictly lower triangular) and whose last rows are the rows of the next generator matrix G_{i+1}, i.e.,*

$$\begin{bmatrix} \mathbf{0} \\ G_{i+1} \end{bmatrix} = \begin{bmatrix} \bar{G}_i \begin{bmatrix} I & \\ & 0 \end{bmatrix} & F_i \bar{l}_i \end{bmatrix}. \tag{3.10}$$

4. *The case $g_i J g_i^T = 0$ is ruled out by the strong regularity of M.*

Schematically, for the special case $r = 2$, we have the following simple array picture for a positive-step case (a similar picture holds for a negative-step case):

$$G_i = \begin{bmatrix} x & x \\ x & x \\ x & x \\ \vdots & \vdots \end{bmatrix} \xrightarrow{\Theta_i} \underbrace{\begin{bmatrix} x' & 0 \\ x' & x' \\ x' & x' \\ \vdots & \vdots \end{bmatrix}}_{\bar{G}_i} \xrightarrow{\text{apply } F_i} \begin{bmatrix} 0 & 0 \\ x'' & x' \\ x'' & x' \\ \vdots & \vdots \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ G_{i+1} & \end{bmatrix}. \tag{3.11}$$

Using words we have the following:

- Use the top row of G_i to define a J-unitary matrix Θ_i that transforms this row to the form $\begin{bmatrix} x' & 0 \end{bmatrix}$;
- multiply G_i by Θ_i and keep the last columns unchanged;
- apply F_i to the first column of $\bar{G}_i = G_i \Theta_i$;
- these two operations result in G_{i+1}.

The rotations Θ_i are always guaranteed to exist and they can be constructed in different ways (see, e.g., [15, Lem. 4.3 and sect. 4.4.1]).

After n steps, the algorithm provides the triangular decomposition [15]

$$M = \sum_{i=1}^{n} d_i l_i l_i^T \tag{3.12}$$

at $O(rn^2)$ computational cost.

Moreover, the successive matrices G_i that are obtained via the algorithm have an interesting interpretation. Let M_i denote the Schur complement of M with respect to its leading $(i-1) \times (i-1)$ submatrix. That is, $M_1 = M$, M_2 is the Schur complement with respect to the $(1,1)$ top left entry of M, M_3 is the Schur complement with respect to the 2×2 top left submatrix of M, and so on. The matrices M_i are therefore $(n-i+1) \times (n-i+1)$. Recall also that F_i denotes the submatrix obtained by deleting the first $(i-1)$ rows and columns of F. Hence, M_i and F_i have the same dimensions.

While the M_i are never computed explicitly, it can be shown that (M_i, F_i, G_i) satisfy the displacement equation [15]

$$M_i - F_i M_i F_i^T = G_i J G_i^T. \tag{3.13}$$

Hence, G_i constitutes a generator matrix for the ith Schur complement M_i, which is therefore structured. Note further that $\bar{G}_i$ is also a generator matrix for the same Schur complement M_i since, due to the J-unitarity of Θ_i, we have $\bar{G}_i J \bar{G}_i^T = G_i \Theta_i J \Theta_i^T G_i^T = G_i J G_i^T$.

We summarize the above discussion in the following statement, deliberately stated in loose terms.

LEMMA 3.2. *The successive Schur complements of a structured matrix are also structured and the generalized Schur algorithm is a recursive procedure that provides generator matrices for the successive Schur complements. It also provides the triangular factors of the original matrix.*

We also indicate here, for later reference, that two successive Schur complements M_i and M_{i+1} are related via the Schur complementation step:

$$M_i = d_i \bar{l}_i \bar{l}_i^T + \begin{bmatrix} 0 & 0 \\ 0 & M_{i+1} \end{bmatrix}. \tag{3.14}$$

We now address the main issues of this paper.

4. Fast QR factorization of shift structured matrices. Let T be an $n \times n$ shift structured matrix (possibly nonsymmetric) with displacement rank r,

$$T - ZTZ^T = GB^T. \tag{4.1}$$

Special cases include the Toeplitz matrix of (2.10) and quasi-Toeplitz matrices of (2.11), whose displacement ranks are equal to 2 ($r = 2$).

Consider the $3n \times 3n$ augmented matrix

$$M = \begin{bmatrix} -I & T & \mathbf{0} \\ T^T & \mathbf{0} & T^T \\ \mathbf{0} & T & \mathbf{0} \end{bmatrix}. \tag{4.2}$$

The matrix M is also structured (as shown below) with respect to $Z_n \oplus Z_n \oplus Z_n$, where Z_n denotes the $n \times n$ lower shift triangular matrix (denoted earlier by Z; here we include the subscript n in order to explicitly indicate the size of Z).

It can be easily verified that $M - (Z_n \oplus Z_n \oplus Z_n)M(Z_n \oplus Z_n \oplus Z_n)^T$ is low rank since

$$M - (Z_n \oplus Z_n \oplus Z_n)M(Z_n \oplus Z_n \oplus Z_n)^T = \begin{bmatrix} -e_1 e_1^T & GB^T & \mathbf{0} \\ BG^T & \mathbf{0} & BG^T \\ \mathbf{0} & GB^T & \mathbf{0} \end{bmatrix}, \tag{4.3}$$

where $e_1 = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}^T$ is a basis vector of appropriate dimension. A generator matrix for M, with $3n$ rows and $(2r+1)$ columns, can be seen to be

$$\mathcal{G} = \frac{1}{\sqrt{2}} \begin{bmatrix} G & -G & e_1 \\ B & B & \mathbf{0} \\ G & -G & \mathbf{0} \end{bmatrix}, \quad \mathcal{J} = \begin{bmatrix} I_r & \\ & -I_{r+1} \end{bmatrix}. \tag{4.4}$$

That is,

$$M - \mathcal{F}M\mathcal{F}^T = \mathcal{G}\mathcal{J}\mathcal{G}^T,$$

where $\mathcal{F} = (Z_n \oplus Z_n \oplus Z_n)$ and $(\mathcal{G}, \mathcal{J})$ are as above.

The $n \times n$ leading submatrix of M is negative definite (in fact, equal to $-I$). Therefore, the first n steps of the generalized Schur algorithm applied to $(\mathcal{F}, \mathcal{G}, \mathcal{J})$ will be negative steps (cf. step 3 of Algorithm 3.1). These first n steps lead to a generator matrix, denoted by $\mathcal{G}_{n+1}$ (with $2n$ rows), for the Schur complement of M with respect to its leading $n \times n$ leading submatrix, viz.,

$$M_{n+1} - (Z_n \oplus Z_n)M_{n+1}(Z_n \oplus Z_n)^T = \mathcal{G}_{n+1}\mathcal{J}\mathcal{G}_{n+1}^T, \tag{4.5}$$

where M_{n+1} is $2n \times 2n$ and equal to

$$M_{n+1} = \begin{bmatrix} T^T T & T^T \\ T & \mathbf{0} \end{bmatrix}. \tag{4.6}$$

Clearly, M and its Schur complement M_{n+1} are related via the Schur complement relation (cf. (3.14))

$$M = \begin{bmatrix} I \\ -T^T \\ \mathbf{0} \end{bmatrix} (-I) \begin{bmatrix} I & -T^T & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & T^T T & T^T \\ \mathbf{0} & T & \mathbf{0} \end{bmatrix}.$$

Therefore, $(\mathcal{G}_{n+1}, \mathcal{J})$ is a generator for M_{n+1} with respect to $(Z_n \oplus Z_n)$, as shown by (4.5).

The leading $n \times n$ submatrix of M_{n+1} is now positive definite (equal to $T^T T$). Therefore, the next n steps of the generalized Schur algorithm applied to $(Z_n \oplus Z_n, \mathcal{G}_{n+1}, \mathcal{J})$ will be positive steps (cf. step 2 of Algorithm 3.1). These steps lead to a generator matrix, denoted by $\mathcal{G}_{2n+1}$ (with n rows), for the Schur complement of M with respect to its leading $2n \times 2n$ leading submatrix, viz.,

$$M_{2n+1} - Z_n M_{2n+1} Z_n^T = \mathcal{G}_{2n+1} \mathcal{J} \mathcal{G}_{2n+1}^T,$$

where M_{2n+1} is now $n \times n$ and equal to $-I$.

Again, M_{n+1} and M_{2n+1} are related via a (block) Schur complementation step (cf. (3.14)), written as

$$\begin{bmatrix} T^T T & T^T \\ T & \mathbf{0} \end{bmatrix} = M_{n+1} = \begin{bmatrix} R^T \\ Q \end{bmatrix} (I) \begin{bmatrix} R & Q^T \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -I \end{bmatrix}, \tag{4.7}$$

where we have denoted the first n columns of the triangular factor of M_{n+1} by

$$\begin{bmatrix} R^T \\ Q \end{bmatrix}$$

with R an $n \times n$ upper triangular matrix and Q an $n \times n$ matrix. The R and Q matrices are thus obtained by splitting the first n columns of the triangular factor of M_{n+1} into a leading lower triangular block followed by a full matrix Q.

By equating terms on both sides of (4.7) we can explicitly identify R and Q as follows:

$$T^T T = R^T R \ , \ \ T = QR \, , \quad QQ^T - I = \mathbf{0}.$$

These relations show that Q and R define the QR factors of the matrix T.

In summary, the above discussion shows the following: given a shift structured matrix T as in (4.1), its QR factorization can be computed efficiently by applying $2n$ steps of the generalized Schur algorithm to the matrices $(\mathcal{F}, \mathcal{G}, \mathcal{J})$ defined in (4.4). The factors Q and R can be obtained from the triangular factors $\{l_i\}$ for $i = n + 1, n + 2, \ldots, 2n$.

Alternatively, if a generator matrix is directly available for M_{n+1} in (4.6) (see section 4.1), then we need only apply n Schur steps to the generator matrix and read the factors Q and R from the resulting n columns of the triangular factor.

In the later sections of this paper we shall establish, for convenience of exposition, the numerical stability of a fast solver for $Tx = b$ that starts with a generator matrix

for the embedding (4.6) rather than the embedding (4.2). It will become clear, however, that the same conclusions will hold if we instead start with a generator matrix for the embedding (4.2).

The augmentation (4.2) was used in [16, 17] and it is based on embedding ideas originally pursued in [5, 13] (see section 4.2).

4.1. The Toeplitz case. In some cases it is possible to find an explicit generator matrix for M_{n+1}. This saves the first n steps of the generalized Schur algorithm.

For example, consider the case when T is a Toeplitz matrix (which is a special case of (4.1) whose first column is $[t_0, t_1, \ldots, t_{n-1}]^T$ and whose first row is $[t_0, t_{-1}, \ldots, t_{-n+1}]$). Define the vectors

$$\begin{bmatrix} c_0 \\ \vdots \\ c_{n-1} \end{bmatrix} = \frac{Te_1}{\|Te_1\|}, \qquad \begin{bmatrix} s_0 \\ \vdots \\ s_{n-1} \end{bmatrix} = T^T \begin{bmatrix} c_0 \\ \vdots \\ c_{n-1} \end{bmatrix}.$$

It can be verified that a generator matrix for M_{n+1} in (4.6) is the following [5]:

$$M_{n+1} - (Z_n \oplus Z_n) M_{n+1} (Z_n \oplus Z_n)^T = \mathcal{G}_{n+1} \mathcal{J} \mathcal{G}_{n+1}^T,$$

where $\mathcal{J}$ is 5×5,

$$\mathcal{J} = \mathrm{diag}[1, 1, -1, -1, -1],$$

and $\mathcal{G}_{n+1}$ is $2n \times 5$,

$$\mathcal{G}_{n+1} = \begin{bmatrix} s_0 & 0 & 0 & 0 & 0 \\ s_1 & t_{-1} & s_1 & t_{n-1} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{n-1} & t_{-n+1} & s_{n-1} & t_1 & 0 \\ c_0 & 1 & c_0 & 0 & 1 \\ c_1 & 0 & c_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ c_{n-1} & 0 & c_{n-1} & 0 & 0 \end{bmatrix}.$$

4.2. Other augmentations. It is also possible to compute the QR factors of a structured matrix T satisfying (4.1) by using other augmented matrices, other than (4.2). For example, consider the $3n \times 3n$ augmented matrix

$$M = \begin{bmatrix} -I & T & \mathbf{0} \\ T^T & \mathbf{0} & T^T \\ \mathbf{0} & T & I \end{bmatrix}, \tag{4.8}$$

where an identity matrix replaces the zero matrix in the $(3,3)$ block entry of the matrix in (4.2). A generator matrix for M, with $3n$ rows and $(2r+2)$ columns, is now

$$\mathcal{G} = \frac{1}{\sqrt{2}} \begin{bmatrix} G & \mathbf{0} & -G & e_1 \\ B & \mathbf{0} & B & \mathbf{0} \\ G & e_1 & -G & \mathbf{0} \end{bmatrix}, \quad \mathcal{J} = \begin{bmatrix} I_{r+1} & \\ & -I_{r+1} \end{bmatrix}.$$

If T is Toeplitz, as in section 4.1, then the rank of $\mathcal{G}$ can be shown to reduce to $2r = 4$ [5] (this is in contrast to the displacement rank 5 that follows from the earlier embedding (4.2), as shown in section 4.1).

After $2n$ steps of the generalized Schur algorithm applied to the above $(\mathcal{G}, \mathcal{J})$, we obtain the following factorization (since now $M_{2n+1} = \mathbf{0}$):

$$M = \begin{bmatrix} I & \mathbf{0} \\ -T^T & R^T \\ \mathbf{0} & Q \end{bmatrix} \begin{bmatrix} -I & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -T^T & R^T \\ \mathbf{0} & Q \end{bmatrix}^T,$$

from which we can again read the QR factors of T from the triangular factors $\{l_i\}$ for $i = n+1, \ldots, 2n+1$. This augmentation was suggested in [5, p. 37] and [13].

However, from a numerical point of view, computing the QR factors of a structured matrix T using the generalized Schur algorithm on the augmented matrices M in (4.2) or (4.8) is not stable. The problem is that the computed Q matrix is not necessarily orthogonal. This is also true for other procedures for fast QR factorization [1, 7, 8, 19].

In the next section we show how to overcome this difficulty and develop a fast and stable algorithm for solving linear systems of equations with shift structured coefficient matrices T. For this purpose, we proceed with the embedding suggested earlier in (4.2) since it seems difficult to obtain a stable algorithm that is based solely on the alternative embedding (4.8). The reason is that the embedding (4.2) allows us to incorporate a correction procedure into the algorithm in order to ensure stability.

We first derive a stable algorithm for a well-conditioned coefficient matrix, and then modify it for the case when the coefficient matrix is ill conditioned. The interested reader may consult at this time the summary of the final algorithm that is provided in section 10.

5. Well-conditioned T. In this section we develop a stable algorithm for the case of well-conditioned matrices T. A definition of what we mean by a well-conditioned matrix is given further ahead (see (5.19)). Essentially this refers to matrices whose condition number is less than the reciprocal of the square root of the machine precision. Modifications to handle the ill-conditioned case will be introduced later in the paper.

We start with an $n \times n$ (possibly nonsymmetric) shift structured matrix T with displacement rank r,

$$T - Z_n T Z_n^T = GB^T, \tag{5.1}$$

and assume we have available a generator matrix $\mathcal{G}$ for the $2n \times 2n$ augmented matrix

$$M = \begin{bmatrix} T^T T & T^T \\ T & \mathbf{0} \end{bmatrix}, \tag{5.2}$$

that is,

$$M - \mathcal{F} M \mathcal{F}^T = \mathcal{G} \mathcal{J} \mathcal{G}^T, \tag{5.3}$$

where $\mathcal{F} = (Z_n \oplus Z_n)$. Note that, for ease of exposition, we have modified our notation. While we have earlier denoted the above matrix M by M_{n+1}, its generator by $\mathcal{G}_{n+1}$, and have used $\mathcal{F}$ to denote $(Z_n \oplus Z_n \oplus Z_n)$, we are now dropping the subscript $n+1$ from $(M_{n+1}, \mathcal{G}_{n+1})$ and are using $\mathcal{F}$ to denote the $2n \times 2n$ matrix $(Z_n \oplus Z_n)$.

In section 4.1 we have discussed an example where we have shown a particular generator matrix $\mathcal{G}$ for M when T is Toeplitz. (We repeat that the error analysis of

later sections will still apply if we instead start with the $3n \times 3n$ embedding (4.2) and its generator matrix (4.4).)

We have indicated earlier (at the end of section 4) that by applying n steps of the generalized Schur algorithm to the matrix M in (5.2) we can obtain the QR factorization of T from the resulting n columns of the triangular factors of M. But this procedure is not numerically stable since the resulting Q is not guaranteed to be unitary. To fix this problem, we propose some modifications. The most relevant modification we introduce now is to run the Schur algorithm for $2n$ steps on M rather than just n steps. As suggested in the paper [4], we also need to be careful in the application of the hyperbolic rotations. In particular, we assume that the hyperbolic rotations are applied using one of the methods suggested in the paper [4] (mixed downdating, OD method, or H procedure; see Appendices A and B at the end of this paper).

The matrix T is only required to be invertible. In this case, the leading submatrix of M in (5.2) is positive definite and therefore the first n steps of the generalized Schur algorithm will be positive steps. Hence, the hyperbolic rotations needed for the first n steps will perform transformations of the form (3.3), where generators are transformed into proper form with respect to their first column. Likewise, the Schur complement of M with respect to its leading submatrix $T^T T$ is equal to $-I$, which is negative definite. This means that the last n steps of the generalized Schur algorithm will be negative steps. Hence, the hyperbolic rotations needed for the last n steps will perform transformations of the form (3.7), where generators are transformed into proper form with respect to their last column.

During a positive step (a similar discussion holds for a negative step), a generator matrix G_i will be reduced to proper form by implementing the hyperbolic transformation Θ_i as a sequence of orthogonal transformations followed by a 2×2 hyperbolic rotation (see also [18]). The 2×2 rotation is implemented along the lines of [4], e.g., via mixed downdating [3], or the OD method, or the H procedure (see Appendices A and B for a description of the OD and H procedures [4]). Details are given below.

5.1. Implementation of the $\mathcal{J}$-unitary rotations Θ_i. When the generalized Schur algorithm is applied to $(\mathcal{G}, \mathcal{F})$ in (5.3), we proceed through a sequence of generator matrices $(\mathcal{G}, \mathcal{G}_2, \mathcal{G}_3, \ldots)$ of decreasing number of rows $(2n, 2n-1, 2n-2, \ldots)$. Let g_i denote the top row of the generator matrix $\mathcal{G}_i$ at step i. In a positive step, it needs to be reduced to the form (3.3) via an $(I_p \oplus -I_q)$-unitary rotation Θ_i. We propose to perform this transformation as follows:

1. Apply a *unitary* (orthogonal) rotation (e.g., Householder) to the first p columns of $\mathcal{G}_i$ so as to reduce the top row of these p columns into proper form,

$$g_i = \begin{bmatrix} x & x & x & x & x & x \end{bmatrix} \overset{unitary\ \Theta_{i,1}}{\longrightarrow} \begin{bmatrix} x & 0 & 0 & x & x & x \end{bmatrix} = g_{i,1},$$

   with a nonzero entry in the first column. Let

$$\mathcal{G}_{i,1} = \mathcal{G}_i \begin{bmatrix} \Theta_{i,1} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \tag{5.4}$$

   denote the modified generator matrix. Its last q columns coincide with those of $\mathcal{G}_i$.
2. Apply another *unitary* (orthogonal) rotation (e.g., Householder) to the last q columns of $\mathcal{G}_{i,1}$ so as to reduce the top row of these last q columns into proper

form with respect to their last column,

$$g_{i,1} = \begin{bmatrix} x & 0 & 0 & x & x & x \end{bmatrix} \overset{unitary\ \Theta_{i,2}}{\longrightarrow} \begin{bmatrix} x & 0 & 0 & 0 & 0 & x \end{bmatrix} = g_{i,2},$$

with a nonzero entry in the last column. Let

$$\mathcal{G}_{i,2} = \mathcal{G}_{i,1} \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & \Theta_{i,2} \end{bmatrix} \tag{5.5}$$

denote the modified generator matrix. Its first p columns coincide with those of $\mathcal{G}_{i,1}$.

3. Employ an elementary hyperbolic rotation $\Theta_{i,3}$ acting on the first and last columns (in mixed-downdating [3] form, or according to the OD or the H methods of [4]; see also Appendices A and B) in order to annihilate the nonzero entry in the last column,

$$g_{i,2} = \begin{bmatrix} x & 0 & 0 & 0 & x \end{bmatrix} \overset{hyperbolic\ \Theta_{i,3}}{\longrightarrow} \begin{bmatrix} x & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

4. The combined effect of the above steps is to reduce g_i to the proper form (3.3) and, hence,

$$\bar{\mathcal{G}}_i = \mathcal{G}_i \begin{bmatrix} \Theta_{i,1} & \mathbf{0} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & \Theta_{i,2} \end{bmatrix} \Theta_{i,3}. \tag{5.6}$$

Expression (5.6) shows that, in infinite precision, the generator matrices $\mathcal{G}_i$ and $\bar{\mathcal{G}}_i$ must satisfy the fundamental requirement

$$\mathcal{G}_i \mathcal{J} \mathcal{G}_i^T = \bar{\mathcal{G}}_i \mathcal{J} \bar{\mathcal{G}}_i^T. \tag{5.7}$$

Obviously, this condition cannot be guaranteed in finite precision. But with the above implementation of the transformation (5.6) (as a sequence of two orthogonal transformations and a hyperbolic rotation in mixed, OD, or H forms), equality (5.7) can be guaranteed to within a "small" error (see (5.8)). Indeed, it follows from (5.4) and (5.5), and from the orthogonality of $\Theta_{i,1}$ and $\Theta_{i,2}$, that

$$\|\hat{\mathcal{G}}_{i,2} \mathcal{J} \hat{\mathcal{G}}_{i,2}^T - \mathcal{G}_i \mathcal{J} \mathcal{G}_i^T\| \leq c_2 \epsilon \|\mathcal{G}_i\|^2,$$

and

$$\left| \|\hat{\mathcal{G}}_{i,2}\|^2 - \|\mathcal{G}_i\|^2 \right| \leq c_3 \epsilon \|\mathcal{G}_i\|^2.$$

It further follows from the error bound (A.3) (in the Appendix) that

$$\|\hat{\bar{\mathcal{G}}}_i \mathcal{J} \hat{\bar{\mathcal{G}}}_i^T - \hat{\mathcal{G}}_{i,2} \mathcal{J} \hat{\mathcal{G}}_{i,2}^T\| \leq c_4 \epsilon \left(\|\hat{\bar{\mathcal{G}}}_i\|^2 + \|\hat{\mathcal{G}}_{i,2}\|^2 \right).$$

Combining the above error bounds we conclude that the following holds:

$$\|\hat{\bar{\mathcal{G}}}_i \mathcal{J} \hat{\bar{\mathcal{G}}}_i^T - \mathcal{G}_i \mathcal{J} \mathcal{G}_i^T\| \leq c_5 \epsilon \left(\|\hat{\bar{\mathcal{G}}}_i\|^2 + \|\mathcal{G}_i\|^2 \right). \tag{5.8}$$

A similar analysis holds for a negative step, where the hyperbolic rotation $\Theta_{i,3}$ is again implemented as a sequence of two unitary rotations and one elementary

hyperbolic rotation in order to guarantee the transformation (3.7). We forgo the details here.

We finally remark that in the algorithm, the incoming generator matrix $\mathcal{G}_i$ will in fact be the computed version, which we denote by $\hat{\mathcal{G}}_i$. This explains why in the error analysis of the next section (see (5.11) and (5.13)) we replace $\mathcal{G}_i$ by $\hat{\mathcal{G}}_i$ in the error bound (5.8).

Note that we are implicitly assuming that the required hyperbolic rotation $\Theta_{i,3}$ exists. While that can be guaranteed in infinite precision, it is possible that in finite precision we can experience breakdowns. This matter is handled in section 5.3.

5.2. Error analysis of the first n steps. After the first n steps of the generalized Schur algorithm applied to $(\mathcal{F}, \mathcal{G})$ in (5.3), we let

$$\begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix}$$

denote the computed factors that correspond to expression (4.7). We further define the matrix S_{n+1} that solves the displacement equation

$$S_{n+1} - Z_n S_{n+1} Z_n^T = \hat{\mathcal{G}}_{n+1} \mathcal{J} \hat{\mathcal{G}}_{n+1}^T. \tag{5.9}$$

Note that S_{n+1} is an $n \times n$ matrix, which in infinite precision would have been equal to the Schur complement $-I$ (cf. (4.7)). We can now define

$$\hat{M} = \begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix} \begin{bmatrix} \hat{R} & \hat{Q}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & S_{n+1} \end{bmatrix}. \tag{5.10}$$

We also define the difference

$$N_i = \hat{\tilde{\mathcal{G}}}_i \mathcal{J} \hat{\tilde{\mathcal{G}}}_i^T \; - \; \hat{\mathcal{G}}_i \mathcal{J} \hat{\mathcal{G}}_i^T, \tag{5.11}$$

and introduce the error matrix $E = M - \hat{M}$. Using (5.8), the error analysis in [4, sect. 7, eq. (41)] can be extended to show that the $2n \times 2n$ error matrix satisfies the equation

$$E - \mathcal{F} E \mathcal{F}^T = \sum_{i=1}^{n} N_i.$$

Consequently, since $\mathcal{F} = (Z_n \oplus Z_n)$ is nilpotent,

$$E = \sum_{k=0}^{n-1} \mathcal{F}^k \left(\sum_{i=1}^{n} N_i \right) (\mathcal{F}^k)^T.$$

If we further invoke the fact that $\mathcal{F}$ is contractive we conclude that

$$\|E\| \leq \sum_{k=0}^{n-1} \left\| \sum_{i=1}^{n} N_i \right\| \; \leq \; \sum_{k=0}^{n-1} \sum_{i=1}^{n} \|N_i\| \; = \; n \sum_{i=1}^{n} \|N_i\|, \tag{5.12}$$

where, according to (5.8),

$$\|N_i\| \; \leq \; c_5 \epsilon \left(\|\hat{\tilde{\mathcal{G}}}_i\|^2 + \|\hat{\mathcal{G}}_i\|^2 \right). \tag{5.13}$$

But since all columns of $\hat{\mathcal{G}}_{i+1}$ and $\hat{\bar{\mathcal{G}}}_i$ coincide, except for one column in $\hat{\bar{\mathcal{G}}}_i$ that is shifted down (multiplied by $\mathcal{F}_i$) to produce the corresponding column in $\hat{\mathcal{G}}_{i+1}$, then we clearly have

$$\|\hat{\mathcal{G}}_{i+1}\|^2 \;\leq\; \|\hat{\bar{\mathcal{G}}}_i\|^2.$$

We can therefore rewrite (5.13) as

$$\|N_i\| \;\leq\; c_6\epsilon\left(\|\hat{\mathcal{G}}_{i+1}\|^2 + \|\hat{\mathcal{G}}_i\|^2\right). \tag{5.14}$$

Substituting into (5.12) we obtain the following error bound:

$$\|E\| \;\leq\; c_7\epsilon\sum_{i=1}^{n}\left(\|\hat{\mathcal{G}}_{i+1}\|^2 + \|\hat{\mathcal{G}}_i\|^2\right) \;\leq\; c_8\epsilon\sum_{i=1}^{n}\|\hat{\mathcal{G}}_i\|^2. \tag{5.15}$$

5.3. Avoiding breakdown. The above error analysis assumes that the first n steps of the generalized Schur algorithm applied to $(\mathcal{G},\mathcal{F})$ in (5.3) do not break down. That is, during the first n steps, the $\mathcal{J}$-unitary rotations Θ_i are well defined. This further requires that the leading submatrices of the first n successive Schur complements remain positive definite. We now show that this can be guaranteed by imposing a lower bound on the minimum singular value of the matrix T (see (5.19); this corresponds to requiring a well-conditioned T, an assumption that will be dropped in section 7 when the algorithm is extended for ill-conditioned T).

The argument is inductive. We assume that the algorithm has successfully completed the first $(i-1)$ steps and define the matrix S_i that solves the displacement equation

$$S_i - \mathcal{F}_i S_i \mathcal{F}_i^T = \hat{\mathcal{G}}_i \mathcal{J} \hat{\mathcal{G}}_i^T, \qquad 1 \leq i \leq (n+1), \tag{5.16}$$

where $\mathcal{F}_i$ is the submatrix obtained from $\mathcal{F}$ in (5.3) by deleting its first $(i-1)$ rows and columns. In particular, $\mathcal{F}_1 = \mathcal{F}$ and $\mathcal{F}_n = Z_n$. Note that S_i is an $(2n-i+1)\times(2n-i+1)$ matrix, which in infinite precision would have been equal to the Schur complement of M with respect to its leading $(i-1)\times(i-1)$ submatrix.

We further define, for $1 \leq i \leq n+1$, the matrices $\hat{M}_i$,

$$\hat{M}_i = \sum_{j=1}^{i-1}\hat{l}_i\hat{l}_i^T \;+\; S_i, \tag{5.17}$$

where the $\hat{l}_i$ are the computed triangular factors, given by (cf. (3.4) and (3.5)). We can again establish, by following the arguments of [4, sect. 7.1], that the error matrices $E_i = M - \hat{M}_i$ satisfy

$$E_i - \mathcal{F}_i E_i \mathcal{F}_i^T = \sum_{j=1}^{i-1} N_j.$$

This relation again establishes, along the lines of (5.15), that

$$\|M - \hat{M}_i\| \;\leq\; c_9\epsilon\sum_{j=1}^{i-1}\|\hat{\mathcal{G}}_j\|^2.$$

Therefore, if the minimum eigenvalue of the leading $n \times n$ submatrix of M (which is equal to $T^T T$) meets the lower bound

$$\lambda_{\min}(T^T T) \; > \; c_9 \epsilon \sum_{j=1}^{i-1} \|\hat{\mathcal{G}}_j\|^2, \tag{5.18}$$

then the leading $n \times n$ submatrix of $\hat{M}_i$ will be guaranteed to be positive definite and the algorithm can continue to the next iteration.

This analysis suggests the following lower bound on the minimum singular value of T in order to avoid breakdown in the first n steps of the algorithm:

$$\sigma^2_{\min}(T) \; > \; 2\, c_9 \epsilon \sum_{j=1}^{n} \|\hat{\mathcal{G}}_j\|^2. \tag{5.19}$$

We refer to a matrix T that satisfies the above requirement as being well conditioned (the scalar multiple 2 is made explicit for convenience in later discussion; see (5.29)).

THEOREM 5.1 (error bound). *The first n steps of the generalized Schur algorithm applied to* $(\mathcal{F}, \mathcal{G})$ *in* (5.3)*, for a matrix* T *satisfying* (5.19)*, and with the rotations* Θ_i *implemented as discussed in section* 5.1*, guarantee the following error bound on the matrix* $(M - \hat{M})$ *(with* $\hat{M}$ *defined in* (5.10)*):*

$$\|M - \hat{M}\| \leq c_9 \epsilon \sum_{j=1}^{n} \|\hat{\mathcal{G}}_j\|^2. \tag{5.20}$$

5.4. Growth of generators. The natural question then is, How big can the norm of the generator matrices be? The analysis that follows is motivated by an observation in [18] that for matrices of the form $T^T T$, with T Toeplitz, there is no appreciable generator growth.

To establish an upper bound on the generator norm, we consider the generator matrix $\hat{\mathcal{G}}_i$ (at the ith step) and recall from the discussion that led to (5.6) that, in a positive step, $\hat{\mathcal{G}}_i$ is transformed via three rotation steps: a unitary rotation $\Theta_{i,1}$ that reduces the first p columns of $\hat{\mathcal{G}}_i$ into proper form, a second unitary rotation $\Theta_{i,2}$ that reduces the last q columns of $\hat{\mathcal{G}}_i$ into proper form, and a last elementary hyperbolic rotation $\Theta_{i,3}$ that reduces the overall generator matrix $\hat{\mathcal{G}}_i$ into proper form.

We denote the first and last columns of $\hat{\mathcal{G}}_i$ by $\hat{u}_i$ and $\hat{v}_i$, respectively, and denote the remaining columns by the block matrices $\hat{U}_i$ and $\hat{V}_i$, i.e., we write

$$\hat{\mathcal{G}}_i = \begin{bmatrix} \hat{u}_i & \hat{U}_i & \hat{V}_i & \hat{v}_i \end{bmatrix}.$$

After the above sequence of three rotations we obtain a new generator matrix $\hat{\tilde{\mathcal{G}}}_i$ that we partition accordingly,

$$\hat{\tilde{\mathcal{G}}}_i = \begin{bmatrix} \hat{\tilde{u}}_i & \hat{\tilde{U}}_i & \hat{\tilde{V}}_i & \hat{\tilde{v}}_i \end{bmatrix}.$$

The last $(r-1)$ columns of $\hat{\tilde{\mathcal{G}}}_i$ remain unchanged and provide the columns of the next generator matrix $\hat{\mathcal{G}}_{i+1}$, while the first column $\hat{\tilde{u}}_i$ is multiplied by $\mathcal{F}_i$ (which essentially corresponds to a simple shifting operation). Hence, we have

$$\hat{\mathcal{G}}_{i+1} = \begin{bmatrix} \hat{u}_{i+1} & \hat{U}_{i+1} & \hat{V}_{i+1} & \hat{v}_{i+1} \end{bmatrix} = \begin{bmatrix} \mathcal{F}_i \hat{\tilde{u}}_i & \hat{\tilde{U}}_i & \hat{\tilde{V}}_i & \hat{\tilde{v}}_i \end{bmatrix}.$$

The first unitary rotation $\Theta_{i,1}$ operates on $\{\hat{u}_i, \hat{U}_i\}$ and provides $\{\tilde{u}_i, \hat{\tilde{U}}_i\}$. This step guarantees the following norm relation:

$$\|\hat{\tilde{U}}_i\| \;\leq\; (1 + c_{10}\epsilon)\left(\|\hat{U}_i\| \;+\; \|\hat{u}_i\|\right).$$

But since $\hat{\tilde{U}}_i = \hat{U}_{i+1}$, we also have

$$\|\hat{U}_{i+1}\| \;\leq\; (1 + c_{10}\epsilon)\left(\|\hat{U}_i\| \;+\; \|\hat{u}_i\|\right).$$

By repeatedly applying the above inequality we obtain

$$\|\hat{U}_{i+1}\| \;\leq\; (1 + c_{11}\epsilon)^i \sum_{j=1}^{i} \|\hat{u}_j\|.$$

Consequently,

$$\left\|\begin{bmatrix} \hat{u}_{i+1} & \hat{U}_{i+1} \end{bmatrix}\right\| \;\leq\; (1 + c_{12}\epsilon)^i \sum_{j=1}^{i+1} \|\hat{u}_j\|. \tag{5.21}$$

But the $\hat{u}_i$, for $i = 2, \ldots, n+1$, are shifted versions of the (nonzero parts of the) columns of the block matrix

$$\begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix}.$$

Therefore,

$$\sum_{j=1}^{i+1} \|\hat{u}_j\| \;\leq\; n \left\| \begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix} \right\| \;+\; \|\hat{u}_1\|.$$

Now further recall that

$$S_{i+1} - \mathcal{F}_{i+1} S_{i+1} \mathcal{F}_{i+1}^T = \hat{\mathcal{G}}_{i+1} \mathcal{J} \hat{\mathcal{G}}_{i+1}^T,$$

where $\mathcal{F}_{i+1}$ is nilpotent (in fact, composed of shift matrices). It thus follows that

$$\left\|\begin{bmatrix} \hat{V}_{i+1} & \hat{v}_{i+1} \end{bmatrix}\right\|^2 \;\leq\; \left\|\begin{bmatrix} \hat{u}_{i+1} & \hat{U}_{i+1} \end{bmatrix}\right\|^2 \;+\; 2\|S_{i+1}\|. \tag{5.22}$$

Combining (5.21) and (5.22) we conclude that

$$\|\hat{\mathcal{G}}_{i+1}\|^2 \;\leq\; 8n^2(1 + c_{12}\epsilon)^{2i} \left\| \begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix} \right\|^2 \;+\; 8\|\hat{u}_1\|^2 \;+\; 4\|S_{i+1}\|. \tag{5.23}$$

We will now show that $\|S_{i+1}\|$ is bounded (at least in infinite precision).

For this purpose, we partition T into $T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}$, where T_1 has i columns and T_2 has $(n-i)$ columns. Commensurately partition M as follows:

$$M = \begin{bmatrix} T_1^T T_1 & T_1^T T_2 & T_1^T \\ T_2^T T_1 & T_2^T T_2 & T_2^T \\ T_1 & T_2 & 0 \end{bmatrix}.$$

Therefore, the Schur complement S_{i+1} in infinite precision is given by

$$S_{i+1} = \begin{bmatrix} T_2^T T_2 - T_2^T T_1 (T_1^T T_1)^{-1} T_1^T T_2 & T_2^T - T_2^T T_1 (T_1^T T_1)^{-1} T_1^T \\ T_2 - T_1 (T_1^T T_1)^{-1} T_1^T T_2 & -T_1 (T_1^T T_1)^{-1} T_1^T \end{bmatrix}.$$

Let the partitioned QR factorization of T in infinite precision be

$$T = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}.$$

Then

$$T_1 (T_1^T T_1)^{-1} T_1^T = Q_1 Q_1^T,$$

which is an orthogonal projector with 2-norm equal to one. It then follows that $\|S_{i+1}\|$ is bounded as follows:

$$\|S_{i+1}\| \ \le \ 1 + 2\|T\|^2 + 2\|T\|. \tag{5.24}$$

The derivation of the above bound can be extended to finite precision by following the technique used in the next section for $\|S_{n+1}\|$. We omit the details here.

Therefore, a first-order bound for the sum of the norms of the generators in (5.20) is given by

$$\begin{aligned} \sum_{i=1}^{n} \|\mathcal{G}_i\|^2 &\le \sum_{i=1}^{n} \left[8n^2 \left\| \begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix} \right\|^2 \ + \ 8\|\hat{u}_1\|^2 \ + \ 4\|S_{i+1}\| \right] \ + \ O(\epsilon^2) \\ &\le 8n^3 \|M\| + 8n\|M\| \ + \ 4n(1 + 2\|T\|^2 + 2\|T\|) \ + \ O(\epsilon^2) \\ &\le 16n(1 + n^2)(1 + \|T\| + \|T^2\|) \ + \ O(\epsilon^2). \end{aligned} \tag{5.25}$$

5.5. Error analysis of the last n steps. It follows from (5.10), and from the definition of $E = M - \hat{M}$, that

$$M - E = \begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix} \begin{bmatrix} \hat{R} & \hat{Q}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & S_{n+1} \end{bmatrix}. \tag{5.26}$$

If we partition the error matrix $-E$ into subblocks, say

$$-E = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}, \qquad E_{21} = E_{12}^T,$$

and use the definition of M in (5.2), we obtain from (5.26) that

$$S_{n+1} = E_{22} - (T + E_{21})(T^T T + E_{11})^{-1}(T^T + E_{12}).$$

Therefore,

$$\begin{aligned} S_{n+1} &= E_{22} - T(T^T T + E_{11})^{-1} T^T - T(T^T T + E_{11})^{-1} E_{12} \\ &\quad - E_{21}(T^T T + E_{11})^{-1} T^T - E_{21}(T^T T + E_{11})^{-1} E_{12} \\ &= -(I + T^{-T} E_{11} T^{-1})^{-1} + \bar{E}, \end{aligned} \tag{5.27}$$

where several terms have been collected into the matrix $\bar{E}$,

$$\bar{E} = E_{22} - T(T^T T + E_{11})^{-1} E_{12} - E_{21}(T^T T + E_{11})^{-1} T^T - E_{21}(T^T T + E_{11})^{-1} E_{12}.$$

Its norm satisfies the bound

$$\|\bar{E}\| \;\leq\; \|E\| + \;\frac{2\,\|T\|\,\|E\|}{\lambda_{\min}(T^TT) - \|E\|} \;+ \frac{\|E\|^2}{\lambda_{\min}(T^TT) - \|E\|},$$

and the denominator is positive in view of (5.19) and (5.20). At this stage we make the following normalization assumption:

$$\|T\| \leq \frac{1}{5}, \tag{5.28}$$

which can always be guaranteed by proper scaling (as explained in the statement of the algorithm in section 10).

We also recall that the well-conditioned assumption (5.19), along with (5.25) and the error bound (5.20), guarantees the following condition:

$$\lambda_{\min}^{-1}(T^TT)\|E\| \leq \frac{1}{2}. \tag{5.29}$$

Remark. This essentially means that the condition number of T should be smaller than $1/\sqrt{\epsilon}$. We will relax this condition in section 7.

From assumptions (5.28) and (5.29) we obtain $\|E\|^2 \leq \|T\|\,\|E\|$, since

$$\|E\| \leq \frac{\sigma_{\min}^2(T)}{2} \;\leq \frac{\|T\|^2}{2} \leq \frac{1}{5}\frac{\|T\|}{2} \leq \|T\|.$$

Therefore,

$$\|\bar{E}\| \;\leq\; \|E\| + \;\frac{3\,\|T\|\,\|E\|}{\lambda_{\min}(T^TT) - \|E\|}.$$

Applying Corollary 8.3.2 in [10] to expression (5.27), we get

$$\sigma_{\min}(S_{n+1}) \geq \frac{1}{1 + \lambda_{\min}^{-1}(T^TT)\|E\|} \;-\; \|\bar{E}\|. \tag{5.30}$$

Using (5.28) and (5.29) we get

$$\sigma_{\min}(S_{n+1}) \geq \frac{2}{3} \;-\; \|\bar{E}\|, \tag{5.31}$$

and

$$\|\bar{E}\| \;\leq\; \frac{11}{5}\|E\| \;\leq\; \frac{11}{25}. \tag{5.32}$$

It then follows from (5.31) that

$$\sigma_{\min}(S_{n+1}) \geq \frac{17}{75} \;\geq \frac{1}{5}. \tag{5.33}$$

We now derive an upper bound for $\|S_{n+1}\|$. Applying Corollary 8.3.2 in [10] to expression (5.27), and using (5.29) and (5.32), we get

$$\sigma_{\max}(S_{n+1}) \leq \frac{1}{1 - \lambda_{\min}^{-1}(T^TT)\|E\|} \;+\; \|\bar{E}\| \leq 2 + \frac{11}{25} \;< 3. \tag{5.34}$$

Therefore, the condition number of S_{n+1} satisfies

$$\kappa(S_{n+1}) \leq 15. \tag{5.35}$$

This establishes that S_{n+1} is a well-conditioned matrix.

By Corollary 8.3.2 in [10], the matrix $(I + T^{-T}E_{11}T^{-1})^{-1}$ in (5.27) is positive definite since by (5.29) $1 - \lambda_{\min}^{-1}(T^TT)\|E\| \geq 1/2 > 0$. Furthermore,

$$\|\bar{E}\| \leq \frac{11}{25} < \frac{1}{2} < \frac{1}{1 - \lambda_{\min}^{-1}(T^TT)\|E\|} \leq 2.$$

Therefore, applying Corollary 8.3.2 in [10] again to expression (5.27) we conclude that S_{n+1} is negative definite.

LEMMA 5.2. *The matrix S_{n+1} defined in* (5.9) *is negative definite and well conditioned. In particular, its condition number is at most* 15 (*cf.* (5.35)).

We can now proceed with the last n steps of the generalized Schur algorithm applied to $\hat{\mathcal{G}}_{n+1}$, since $\hat{\mathcal{G}}_{n+1}$ is a generator matrix for S_{n+1}:

$$S_{n+1} - Z_n S_{n+1} Z_n^T = \hat{\mathcal{G}}_{n+1} \mathcal{J} \hat{\mathcal{G}}_{n+1}^T.$$

All steps will now be negative steps. Hence, the discussion of section 5.1 applies. The only difference will be that we make the generator proper with respect to its last column. In other words, the third step of the algorithm in section 5.1 should be modified as follows:

$$g_{i,2} = \begin{bmatrix} x & 0 & 0 & 0 & x \end{bmatrix} \overset{\text{hyperbolic } \Theta_{i,3}}{\longrightarrow} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & x \end{bmatrix}. \tag{5.36}$$

Let $-\Delta\Delta^T$ be the computed triangular factorization of S_{n+1}. A similar error analysis to that of section 5.2 (or the results of [4]) can be used to show that

$$\|S_{n+1} - (-\Delta\Delta^T)\| \leq c_{13}\epsilon \sum_{i=n+1}^{2n} \|\hat{\mathcal{G}}_i\|^2. \tag{5.37}$$

The norm of the generators $\{\hat{\mathcal{G}}_i\}$ appearing in the above error expression can be shown to be bounded as follows. Similar to (5.21) we have

$$\left\| \begin{bmatrix} \hat{V}_{i+1} & \hat{v}_{i+1} \end{bmatrix} \right\| \leq (1 + c_{14}\epsilon)^{i-n} \sum_{j=n+1}^{i} \|\hat{v}_j\|. \tag{5.38}$$

Moreover, the $\hat{v}_i$, for $i = n+2, \ldots, 2n$, are shifted versions of the (nonzero parts of the) columns of Δ. Hence,

$$\sum_{j=n+1}^{i} \|\hat{v}_j\| \leq n\|\Delta\| + \|\hat{v}_{n+1}\|.$$

By using the fact that Z_n is lower triangular and contractive and that S_{n+1} is negative definite, Lemma B.2 in [4] can be extended to show that

$$\left\| \begin{bmatrix} \hat{u}_{i+1} & \hat{U}_{i+1} \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} \hat{V}_{i+1} & \hat{v}_{i+1} \end{bmatrix} \right\|.$$

Therefore,

$$\begin{aligned}\|\hat{\mathcal{G}}_i\| &\leq \left\|\begin{bmatrix} \hat{u}_{i+1} & \hat{U}_{i+1} \end{bmatrix}\right\| + \left\|\begin{bmatrix} \hat{V}_{i+1} & \hat{v}_{i+1} \end{bmatrix}\right\| \\ &\leq 2\left\|\begin{bmatrix} \hat{V}_{i+1} & \hat{v}_{i+1} \end{bmatrix}\right\| \\ &\leq 2(1+c_{14}\epsilon)^{i-n}[n\|\Delta\| + \|\hat{v}_{n+1}\|], \end{aligned} \tag{5.39}$$

where in infinite precision

$$\|\Delta\|^2 = \|S_{n+1}\| \leq 3,$$

from relation (5.34). Similarly, the bound for $\hat{v}_{n+1}$ follows from (5.23) and (5.24).

Summary. We have shown so far that if we apply $2n$ steps of the generalized Schur algorithm to the matrices $(\mathcal{F}, \mathcal{G})$ in (5.3), with proper implementation of the $\mathcal{J}$-unitary rotations (as explained in section 5.1), then the error in the computed factorization of M is bounded as follows:

$$\left\| M - \begin{bmatrix} \hat{R}^T & 0 \\ \hat{Q} & \Delta \end{bmatrix} \begin{bmatrix} \hat{R} & \hat{Q}^T \\ 0 & -\Delta^T \end{bmatrix} \right\| \leq c_{15}\epsilon \sum_{i=1}^{2n} \|\hat{\mathcal{G}}_i\|^2. \tag{5.40}$$

We have also established (at least in infinite precision) that the norm of the generators is bounded. Therefore, the computed factorization is (at least asymptotically) backward stable with respect to M.

6. Solving linear systems. We now return to the problem of solving the linear system of equations $Tx = b$, where T is a well-conditioned nonsymmetric shift structured matrix (e.g., Toeplitz, quasi-Toeplitz, and product of two Toeplitz matrices).

Note from the bound (5.40) that

$$\|\hat{Q}\hat{Q}^T - \Delta\Delta^T\| \leq c_{15}\epsilon \sum_{i=1}^{2n} \|\hat{\mathcal{G}}_i\|^2.$$

Therefore,

$$\|(\Delta^{-1}\hat{Q})(\Delta^{-1}\hat{Q})^T - I\| \leq c_{15}\epsilon\|\Delta^{-1}\|^2 \sum_{i=1}^{2n} \|\hat{\mathcal{G}}_i\|^2.$$

It follows from (5.33) and (5.35) that

$$\sigma_{\min}(\Delta\Delta^T) \geq \sigma_{\min}(S_{n+1}) - c_{15}\epsilon \sum_{i=n+1}^{2n} \|\hat{\mathcal{G}}_i\|^2 \geq \frac{1}{5} - c_{15}\epsilon \sum_{i=n+1}^{2n} \|\hat{\mathcal{G}}_i\|^2 \approx \frac{1}{5}.$$

Therefore, $\|\Delta^{-1}\|^2$ is bounded by $1/5$ (approximately), from which we can conclude that $\Delta^{-1}\hat{Q}$ is numerically orthogonal.

Furthermore, from (5.40) we also have

$$\|T - \hat{Q}\hat{R}\| \leq c_{15}\epsilon \sum_{i=1}^{2n} \|\hat{\mathcal{G}}_i\|^2.$$

This shows that we can compute x by solving the nearby linear system

$$\Delta\Delta^{-1}\hat{Q}\hat{R}x = b,$$

in $O(n^2)$ flops by exploiting the fact that $\Delta^{-1}\hat{Q}$ is numerically orthogonal and Δ is triangular as follows:

$$\hat{x} \leftarrow \hat{R}^{-1}(\hat{Q}^T\Delta^{-T})\Delta^{-1}b. \tag{6.1}$$

The fact that this scheme for computing x is backward stable will be established in section 8 (see the remark after expression (8.2)).

7. Ill-conditioned T. We now consider modifications to the algorithm when the inequality (5.29) is not satisfied by T. This essentially means that the condition number of T is larger than the square root of the reciprocal of the machine precision. We will refer to such matrices T as being ill conditioned.

There are several potential numerical problems now, all of which have to be eliminated. First, the $(1,1)$ block of M can fail to factorize as it is not sufficiently positive definite. Second, even if the first n steps of the Schur algorithm are completed successfully, the Schur complement S_{n+1} of the $(2,2)$ block may no longer be negative definite, making the algorithm unstable. Third, the matrix Δ may no longer be well conditioned, in which case it is not clear how one can solve the linear system $Tx = b$ in a stable manner. We now show how these problems can be resolved.

To resolve the first two problems we add small multiples of the identity matrix to the $(1,1)$ and $(2,2)$ blocks of M, separately:

$$M = \begin{bmatrix} T^TT + \alpha I & T \\ T^T & -\beta I \end{bmatrix}, \tag{7.1}$$

where α and β are positive numbers that will be specified later.[1] This leads to an increase in the displacement rank of M. For Toeplitz matrices the rank increases only by one and the new generators are given as follows:

$$M - (Z_n \oplus Z_n)M(Z_n \oplus Z_n)^T = \mathcal{G}\mathcal{J}\mathcal{G}^T, \tag{7.2}$$

where $\mathcal{J}$ is 6×6,

$$\mathcal{J} = \mathrm{diag}[1, 1, 1, -1, -1, -1], \tag{7.3}$$

and $\mathcal{G}$ is $2n \times 6$,

$$\mathcal{G} = \begin{bmatrix} \sqrt{\alpha} & s_0 & 0 & 0 & 0 & 0 \\ 0 & s_1 & t_{-1} & s_1 & t_{n-1} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & s_{n-1} & t_{-n+1} & s_{n-1} & t_1 & 0 \\ 0 & c_0 & 1 & c_0 & 0 & \sqrt{1+\beta} \\ 0 & c_1 & 0 & c_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & c_{n-1} & 0 & c_{n-1} & 0 & 0 \end{bmatrix}. \tag{7.4}$$

Had we instead started with the embedding (4.2) for more general shift structured matrices, we would then modify the generators as explained later in the remark in section 9.

[1] We continue to use M for the new matrix in (7.1) for convenience of notation.

Assume α is chosen such that

$$\alpha \;\geq\; c_{16}\epsilon \sum_{j=1}^{n} \|\hat{\mathcal{G}}_j\|^2; \tag{7.5}$$

then since

$$\lambda_{\min}(T^T T + \alpha I) \;>\; c_{16}\epsilon \sum_{j=1}^{n} \|\hat{\mathcal{G}}_j\|^2 \;, \tag{7.6}$$

it follows from the analysis in section 5.3 that the first n steps of the generalized Schur algorithm applied to $\mathcal{G}$ in (7.4) will complete successfully. As in (5.10), define the matrix

$$\hat{M} = \begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix} \begin{bmatrix} \hat{R} & \hat{Q}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & S_{n+1} \end{bmatrix}, \tag{7.7}$$

where S_{n+1} is the solution of

$$S_{n+1} - Z_n S_{n+1} Z_n^T = \hat{\mathcal{G}}_{n+1} \mathcal{J} \hat{\mathcal{G}}_{n+1}.$$

(Recall that $\hat{\mathcal{G}}_{n+1}$ now has six columns and $\mathcal{J}$ is 6×6.) Then following the analysis of the first n steps of section 5.2 we obtain (cf. (5.20))

$$\|E\| = \left\| M - \hat{M} \right\| \;\leq\; c_{19}\epsilon \sum_{j=1}^{n} \|\hat{\mathcal{G}}_j\|^2,$$

where, as shown earlier in (5.23),

$$\|\hat{\mathcal{G}}_{i+1}\|^2 \;\leq\; 8n^2(1 + c_{16}\epsilon)^{2i} \left\| \begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix} \right\|^2 \;+\; 8\|\hat{u}_1\|^2 \;+\; 4\|S_{i+1}\|^2. \tag{7.8}$$

The proof that S_{i+1} is bounded is similar to the proof that S_{n+1} is bounded, which we now give. First, we assume that β satisfies the following bound:

$$\beta \;\geq\; \frac{1 + c_{16}\epsilon}{1 - c_{16}\epsilon}(\|E\| + 4) \;. \tag{7.9}$$

Recall that S_{n+1} satisfies the relation

$$M - E = \begin{bmatrix} \hat{R}^T \\ \hat{Q} \end{bmatrix} \begin{bmatrix} \hat{R} & \hat{Q}^T \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & S_{n+1} \end{bmatrix}. \tag{7.10}$$

If we partition the error matrix $-E$ into subblocks, say

$$-E = \begin{bmatrix} E_{11} & E_{12}^T \\ E_{12} & E_{22} \end{bmatrix},$$

and use the definition of M in (7.1), we obtain from (7.10) that

$$S_{n+1} = -\beta I + E_{22} - (T + E_{12})(T^T T + \alpha I + E_{11})^{-1}(T^T + E_{12}^T). \tag{7.11}$$

Since α and β satisfy (7.5) and (7.9), we have that

$$\alpha \geq \|E\| \geq \|E_{11}\|, \qquad \beta \geq \frac{1+c_{16}\epsilon}{1-c_{16}\epsilon}(\|E\|+4) \geq \|E\| \geq \|E_{22}\|.$$

Therefore, $(\alpha I + E_{11})$ is positive definite and $(-\beta I + E_{22})$ is negative definite. This shows, in view of (7.11), that S_{n+1} is negative definite. We now proceed to bound the smallest and the largest eigenvalues of S_{n+1}.

Using (7.11) we write

$$S_{n+1} = -\beta I + E_{22} - (I + E_{12}T^{-1})(I + \alpha T^{-T}T^{-1} + T^{-T}E_{11}T^{-1})^{-1}(I + T^{-T}E_{12}^T),$$

and note that

$$\|(I + \alpha T^{-T}T^{-1} + T^{-T}E_{11}T^{-1})^{-1}\| = \left\| \left(I + \alpha T^{-T} \left[I + \frac{E_{11}}{\alpha} \right] T^{-1} \right)^{-1} \right\| \leq 1,$$

since $\|E_{11}\|/\alpha < 1$.

We now make the assumption

$$\|T^{-1}\| \, \|E\| \leq 1, \tag{7.12}$$

which is considerably weaker than the assumption (5.29) used in the well-conditioned case. Assumption (7.12) essentially means that the condition number of T should be less than the reciprocal of the machine precision.

It then follows that

$$\|S_{n+1}\| \leq \beta + \|E\| + 4.$$

Since, technically, $\|E\|$ depends upon $\|S_{n+1}\|$, we have only shown that $\|S_{n+1}\|$ is bounded to first order in ϵ. With more effort, this restriction can be removed.

Before proceeding, we mention that the error in factorizing S_{n+1} into $-\Delta\Delta^T$ by the generalized Schur algorithm can be written in the form

$$\|S_{n+1} - (-\Delta\Delta^T)\| \leq c_{17}\epsilon\|S_{n+1}\|,$$

where c_{17} can be obtained by extending the analysis of section 5.2.

As mentioned earlier (cf. (5.18)), S_{n+1} can be factorized by the Schur algorithm if its minimum eigenvalue satisfies

$$|\lambda_{\min}(S_{n+1})| \geq c_{17}\epsilon\|S_{n+1}\| \, .$$

But since $|\lambda_{\min}(S_{n+1})| \geq \beta - \|E_{22}\|$, the above condition can be guaranteed by choosing

$$\begin{aligned}
\beta &\geq c_{17}\epsilon\|S_{n+1}\| + \|E_{22}\| \\
&\geq c_{17}\epsilon(\beta + \|E\| + 4) + \|E\| \\
&\geq \frac{1}{1-c_{17}\epsilon}\left[c_{17}\epsilon(\|E\|+4) + \|E\|\right] \\
&\geq \frac{1+c_{17}\epsilon}{1-c_{17}\epsilon}(\|E\|+4),
\end{aligned}$$

which is assumption (7.9) on β (with $c_{17} = c_{16}$).

Therefore, the last n steps of the generalized Schur algorithm can be completed to give the following error bound in the factorization of M in (7.1):

$$\left\| M - \begin{bmatrix} \hat{R}^T & \mathbf{0} \\ \hat{Q} & \Delta \end{bmatrix} \begin{bmatrix} \hat{R} & \hat{Q}^T \\ \mathbf{0} & -\Delta^T \end{bmatrix} \right\| \leq \alpha + \beta + c_{18}\epsilon \sum_{i=1}^{2n} \|\hat{\mathcal{G}}_i\|^2 , \tag{7.13}$$

where the norm of the generators is again bounded by arguments similar to those in section 5.4. In other words, we have a backward stable factorization of M.

Since Δ is no longer provably well conditioned, we cannot argue that $\Delta^{-1}\hat{Q}$ is numerically orthogonal. For this reason, we now discuss how to solve the linear system of equations $Tx = b$ in the ill-conditioned case.

8. Solving the linear system. Note that if x solves $Tx = b$, then it also satisfies

$$\begin{bmatrix} T^T T & T^T \\ T & \mathbf{0} \end{bmatrix} \begin{bmatrix} x \\ -b \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ b \end{bmatrix}.$$

Using the above backward stable factorization (7.13) we can solve the above linear system of equations to get

$$\left(\begin{bmatrix} T^T T & T^T \\ T & \mathbf{0} \end{bmatrix} + H \right) \begin{bmatrix} \hat{y} \\ \hat{z} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ b \end{bmatrix}, \tag{8.1}$$

where the error matrix H satisfies

$$\|H\| \leq \alpha + \beta + c_{18}\epsilon \sum_{i=1}^{2n} \|\hat{\mathcal{G}}_i\|^2 + c_{19}\epsilon \left\| \begin{bmatrix} \hat{R}^T & \mathbf{0} \\ \hat{Q} & \Delta \end{bmatrix} \right\|^2 .$$

Note that $\hat{y}$ is computed by the expression

$$R^{-1}\hat{Q}^T\Delta^{-T}\Delta^{-1}b, \tag{8.2}$$

which is identical to the earlier formula (6.1) we obtained by assuming $\Delta^{-1}\hat{Q}$ is numerically orthogonal! Therefore, the subsequent error analysis holds equally well for the well-conditioned case.

Moreover, it follows from (8.1) that

$$(T + H_{21})\hat{y} + H_{22}\hat{z} = b.$$

Therefore, we can write this as

$$\left(T + H_{21} + \frac{H_{22}\hat{z}\hat{y}^T}{\hat{y}^T\hat{y}} \right) \hat{y} = b, \tag{8.3}$$

where

$$\left\| H_{21} + \frac{H_{22}\hat{z}\hat{y}^T}{\hat{y}^T\hat{y}} \right\| \leq \|H_{21}\| + \|H_{22}\| \frac{\|\hat{z}\|}{\|\hat{y}\|}. \tag{8.4}$$

If we assume

$$\|T\| \leq 1 \tag{8.5}$$

(which is implied by (5.28)), then in infinite precision

$$\frac{\|\hat{z}\|}{\|\hat{y}\|} = \frac{\|b\|}{\|x\|} = \frac{\|b\|}{\|T^{-1}b\|} \leq \frac{\|b\|}{\|b\|}\,\|T\| \;\leq 1.$$

Under the assumptions in Theorem C.1, which are of a similar nature to assumptions we have already made, we can show that $\|\hat{z}\|/\|\hat{y}\|$ is also bounded in finite precision. Therefore, our algorithm is backward stable for solving shift structured linear systems.

Theorem C.1 imposes a bound on $\kappa(M)$, the condition number of M. We now verify that $\kappa(M)$ is of the same order as $\kappa(T)$. First, note that

$$\|M\| \;\leq\; 2\|T\| \;+\; \|T\|^2 \;\leq\; 3\|T\|,$$

since $\|T\| \leq 1$. Moreover,

$$M^{-1} = \begin{bmatrix} \mathbf{0} & T^{-1} \\ T^{-T} & -I \end{bmatrix},$$

from which we conclude that

$$\|M^{-1}\| \;\leq\; 1 + 2\|T^{-1}\|.$$

Hence,

$$\kappa(M) \;\leq\; (1 + 2\|T^{-1}\|)(3\|T\|) \;\leq\; 9\kappa(T).$$

Therefore, the restriction on $\kappa(M)$ can be considered a restriction on $\kappa(T)$, which will be similar to our earlier assumption (7.12).

For convenience we now give a simple first-order bound for the backward error in (8.3). Indeed,

$$\begin{aligned}
\left\| H_{21} + \frac{H_{22}\hat{z}\hat{y}^T}{\hat{y}^T\hat{y}} \right\| &\leq \|H_{21}\| \;+\; \|H_{22}\| \;+\; O(\epsilon^2) \\
&\leq 2\|H\| \;+\; O(\epsilon^2) \\
&\leq 2\left[\alpha + \beta + c_{20}\epsilon \sum_{i=1}^{2n} \|\hat{\mathcal{G}}_i\|^2 \;+\; c_{21}\epsilon \left\| \begin{bmatrix} \hat{R}^T & \mathbf{0} \\ \hat{Q} & \Delta \end{bmatrix} \right\|^2 \right] \;+\; O(\epsilon^2) \\
&\leq 2(\alpha + \beta) \\
&\quad + c_{22}\epsilon \left[\|M\| \;+\; \sum_{i=1}^{2n} \left(8n^2\|M\| + 4(1 + 2\|T\|^2 + 2\|T\|) \right) \right] \;+\; O(\epsilon^2) \\
&\leq 2(\alpha + \beta) + c_{23}\epsilon \left[\|M\| \;+\; 4n(1 + 2\|T\|^2 + 2\|T\|) \right] \;+\; O(\epsilon^2) \\
&\leq 2(\alpha + \beta) + c_{24}\epsilon[\|M\| + 1] \;+\; O(\epsilon^2) \\
&\leq 2(\alpha + \beta) + c_{25}\epsilon[\|T\| + 1] \;+\; O(\epsilon^2). \qquad (8.6)
\end{aligned}$$

Note that $\|T\|$ should be approximately one for the algorithm to be backward stable. This can be satisfied by appropriately normalizing $\|T\|$.

8.1. Conditions on the coefficient matrix. For ease of reference, we list here the conditions imposed on the coefficient matrix T in order to guarantee a fast backward stable solver of $Tx = b$:

1. $\|T\|$ is suitably normalized to guarantee $\|T\| \approx 1$ (cf. (5.28) and (8.5)).
2. $\|T^{-1}\|$ satisfies (7.12), which essentially means that the condition number of T should be less than the reciprocal of the machine precision.

9. A remark. Had we instead started with the embedding (4.2), we first perform n steps of the generalized Schur algorithm to get a generator matrix $\hat{\mathcal{G}}_{n+1}$ for the computed version of the $2n \times 2n$ embedding (4.6). We then add two columns to $\hat{\mathcal{G}}_{n+1}$ as follows:

$$\begin{bmatrix} \sqrt{\alpha} & & 0 \\ 0 & & 0 \\ 0 & & \sqrt{\beta} \\ \vdots & \hat{\mathcal{G}}_{n+1} & \vdots \\ 0 & & 0 \\ 0 & & 0 \end{bmatrix},$$

where the entry $\sqrt{\beta}$ occurs in the $(n+1)$th row of the last column. The new first column has a positive signature and the new last column has a negative signature.

10. Pseudocode of the algorithm for Toeplitz systems. For convenience we summarize the algorithm here for the case of nonsymmetric Toeplitz systems. We hasten to add though that the algorithm also applies to more general shift structured matrices T (such as quasi Toeplitz or with higher displacement ranks, as demonstrated by the analysis in the earlier sections). The only difference will be in the initial generator matrix $\mathcal{G}$ and signature matrix $\mathcal{J}$ for M in (7.1) and (7.2). The algorithm will also be essentially the same, apart from an additional n Schur steps, if we instead employ the embedding (4.2).

Input: A nonsymmetric $n \times n$ Toeplitz matrix T and an n-dimensional column vector b. The entries of the first column of T are denoted by $[t_0, t_1, \ldots, t_{n-1}]^T$, while the entries of the first row of T are denoted by $[t_0, t_{-1}, \ldots, t_{-n+1}]$.

Output: A backward stable solution of $Tx = b$.

Algorithm:

- Normalize T and b. Since the Frobenius norm of $\|T\|$ is less than

$$\gamma \;=\; \sqrt{n \sum_{i=-n+1}^{n-1} t_i^2},$$

  we can normalize T by setting t_i to be $t_i/(5\gamma)$ for all i. Similarly, divide the entries of b by 5γ. In what follows, T and b will refer to these normalized quantities.
- Define the vectors

$$\begin{bmatrix} c_0 \\ \vdots \\ c_{n-1} \end{bmatrix} = \frac{Te_1}{\|Te_1\|}, \qquad \begin{bmatrix} s_0 \\ \vdots \\ s_{n-1} \end{bmatrix} = T^T \begin{bmatrix} c_0 \\ \vdots \\ c_{n-1} \end{bmatrix}.$$

- Construct the 6×6 signature matrix

$$\mathcal{J} = \mathrm{diag}[1, 1, 1, -1, -1, -1],$$

and the $2n \times 6$ generator matrix $\mathcal{G}$,

$$\mathcal{G} = \begin{bmatrix} \sqrt{\alpha} & s_0 & 0 & 0 & 0 & 0 \\ 0 & s_1 & t_{-1} & s_1 & t_{n-1} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & s_{n-1} & t_{-n+1} & s_{n-1} & t_1 & 0 \\ 0 & c_0 & 1 & c_0 & 0 & \sqrt{1+\beta} \\ 0 & c_1 & 0 & c_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & c_{n-1} & 0 & c_{n-1} & 0 & 0 \end{bmatrix},$$

where the small positive numbers α and β are chosen as follows (by experimental tuning):

$$\alpha = n^{1/2}\epsilon \, \|\mathcal{G}\|^2, \quad \beta = 4(2n)^{1/4}\epsilon.$$

(If T is well conditioned, then we set $\beta = 0 = \alpha$, and delete the first columns of $\mathcal{G}$ and $\mathcal{J}$, which then become $2n \times 5$ and 5×5, respectively).

- Apply n steps of the generalized Schur algorithm starting with $\mathcal{G}_1 = \mathcal{G}$ and $\mathcal{F} = (Z_n \oplus Z_n)$, and ending with $\mathcal{G}_{n+1}$ and $\mathcal{F} = Z_n$. These are positive steps according to the description of Algorithm 3.1 (step 2), where the successive generators are reduced to proper form relative to their first column. Note that this must be performed with care for numerical stability as explained in section 5.1.
- Apply n more steps of the generalized Schur algorithm starting with $\mathcal{G}_{n+1}$. These are negative steps according to the description of Algorithm 3.1 (step 3), where the successive generators are reduced to proper form relative to their last column. This also has to be performed with care as explained prior to (5.36).
- Each of the above $2n$ steps provides a column of the triangular factorization of the matrix M in (7.1), as described in Algorithm 3.1 (steps 2 and 3). The triangular factor of M is then partitioned to yield the matrices $\{\hat{R}, \hat{Q}, \Delta\}$,

$$\begin{bmatrix} \hat{R}^T & \mathbf{0} \\ \hat{Q} & \Delta \end{bmatrix},$$

where $\hat{R}$ is upper triangular and Δ is lower triangular.
- The solution $\hat{x}$ is obtained by evaluating the quantity

$$R^{-1}\hat{Q}^T\Delta^{-T}\Delta^{-1}b,$$

via a sequence of back substitutions and matrix–vector multiplications. The computed solution is backward stable. It satisfies

$$(T + H)\hat{x} = b,$$

where the norm of the error matrix is bounded by

$$(10.1)\quad \|H\| \;\le\; 2(\alpha+\beta) \;+ c_{26}\epsilon[1 + \|T\|] + O(\epsilon^2) \;\le c_{27}\epsilon \, \|T\| + O(\epsilon^2).$$

10.1. Operation count. The major computational cost is due to the application of the successive steps of the generalized Schur algorithm. The overhead operations that are required for the normalization of T, and for the determination of the generator matrix $\mathcal{G}$, amount at most to $O(n \log n)$ flops. Table 10.1 shows the number of flops needed at each step of the algorithm (i denotes the iteration number and it runs from $i = 2n$ down to $i = 1$). The operation count given below assumes that, for each iteration, two Householder transformations are used to implement the reduction to proper form of section 5.1, combined with an elementary hyperbolic rotation in OD form.

TABLE 10.1
Complexity analysis of the algorithm.

During each iteration of the algorithm	Count in flops
Compute two Householder transformations	$3r$
Apply the Householder transformations	$4 \cdot i \cdot r$
Compute the hyperbolic transformation	7
Apply the hyperbolic transformation using OD	$6 \cdot i$
Shift columns	i
Total for $i = 2n$ down to 1	$(14 + 8r)n^2 + 10nr + 21n$
Cost of three back substitution steps	$3n^2$
Cost of matrix–vector multiplication	$2n^2$
Startup costs	$n(24 \log n + r + 52)$
Total cost of the algorithm	$(19 + 8r)n^2 + n(24 \log n + 11r + 73)$

Table 10.2 indicates the specific costs for different classes of structured matrices.

TABLE 10.2
Computational cost for some structured matrices.

Matrix type	Cost
Well-conditioned Toeplitz matrix	$59n^2 + n(24 \log n + 128)$
Ill-conditioned Toeplitz matrix	$67n^2 + n(24 \log n + 139)$

11. Conclusions. We performed extensive experiments to verify the theoretical bounds for both well-conditioned and ill-conditioned Toeplitz matrices. The error was always better than the bounds predicted by the theory. Interested readers can get MATLAB codes of the algorithm by contacting the authors.

The results of this work can be extended to Toeplitz least-squares problems, which will be addressed in a companion paper. Furthermore, there are also useful applications of these ideas in filtering theory, which will be reported elsewhere.

Appendix A. The OD 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,

$$\begin{bmatrix} x_1 & y_1 \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \Theta.$$

The advantage of the OD method is that the computed quantities $\hat{x}_1$ and $\hat{y}_1$ satisfy the equation

$$\text{(A.1)} \qquad \begin{bmatrix} \hat{x}_1 + e_1 & \hat{y}_1 + e_2 \end{bmatrix} = \begin{bmatrix} x + e_3 & y + e_4 \end{bmatrix} \Theta,$$

with

$$\text{(A.2)} \quad \left\| \begin{bmatrix} e_1 & e_2 \end{bmatrix} \right\| \le c_{28}\epsilon \left\| \begin{bmatrix} \hat{x}_1 & \hat{y}_1 \end{bmatrix} \right\|, \quad \left\| \begin{bmatrix} e_3 & e_4 \end{bmatrix} \right\| \le c_{29}\epsilon \left\| \begin{bmatrix} x & y \end{bmatrix} \right\|,$$

and, consequently,

$$\text{(A.3)} \qquad |(\hat{x}_1^2 - \hat{y}_1^2) - (x_1^2 - y_1^2)| \ \le \ c_{30}\epsilon(\hat{x}_1^2 + \hat{y}_1^2 + x^2 + y^2).$$

ALGORITHM A.1 (the OD procedure). *Consider a hyperbolic rotation Θ with reflection coefficient $\rho = \beta/\alpha$, $|\rho| < 1$. Given a row vector $\begin{bmatrix} x & y \end{bmatrix}$ as a prearray, the transformed (postarray) row vector $\begin{bmatrix} x_1 & y_1 \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \Theta$ is computed as follows:*

$$\begin{bmatrix} x' & y' \end{bmatrix} \leftarrow \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

$$\begin{bmatrix} x'' & y'' \end{bmatrix} \leftarrow \begin{bmatrix} x' & y' \end{bmatrix} \begin{bmatrix} \frac{1}{2}\sqrt{\frac{\alpha+\beta}{\alpha-\beta}} & 0 \\ 0 & \frac{1}{2}\sqrt{\frac{\alpha-\beta}{\alpha+\beta}} \end{bmatrix},$$

$$\begin{bmatrix} x_1 & y_1 \end{bmatrix} \leftarrow \begin{bmatrix} x'' & y'' \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

Appendix B. 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 H method is that the computed quantities $\hat{x}_1$ and $\hat{y}_1$ satisfy the equation

$$\text{(B.1)} \qquad \begin{bmatrix} \hat{x}_1 + e_1' & \hat{y}_1 + e_2' \end{bmatrix} = \begin{bmatrix} x & y \end{bmatrix} \Theta,$$

where the error terms satisfy

$$\text{(B.2)} \qquad |e_1'| \le c_{31}\epsilon|\hat{x}_1|, \qquad |e_2'| \le c_{32}\epsilon(|\hat{x}_1| + |\hat{y}_1|).$$

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 $|x| > |y|$.

ALGORITHM B.1 (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
$x_1 \leftarrow \frac{|\alpha| x \xi}{\sqrt{(\alpha-\beta)(\alpha+\beta)}}$
$y_1 \leftarrow x_1 - \sqrt{\frac{\alpha+\beta}{\alpha-\beta}}\,(x-y)$.

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. (B.1)) whereas the OD method is only stable (cf. (A.1)).

Appendix C. Miscellaneous error bounds. The following is an extension of Lemma 2.7.1 and Theorem 2.7.2 of [10].

THEOREM C.1. *Suppose*

$$M \begin{bmatrix} y \\ z \end{bmatrix} = b,$$

where M *is an* $n \times n$ *matrix,* b *is an* n*-dimensional vector, and* $\|z\| \leq \|y\|$. *Let*

$$(M + H) \begin{bmatrix} \hat{y} \\ \hat{z} \end{bmatrix} = b,$$

where H *is an* $n \times n$ *matrix such that* $\|H\| \leq c_{33}\epsilon\|M\|$. *If* $c_{33}\epsilon\kappa(M) = r < \frac{1}{5}$, *where* $\kappa(M) = \|M\|\,\|M^{-1}\|$, *then*

$$\frac{\|\hat{z}\|}{\|\hat{y}\|} \;\leq\; \frac{1+3r}{1-5r}.$$

Proof. From Theorem 2.7.2 in [10] it follows that

$$\|y\| - \frac{2r}{1-r}[\|y\| + \|z\|] \;\leq\; \|\hat{y}\| \;\leq\; \|y\| + \frac{2r}{1-r}[\|y\| + \|z\|].$$

By interchanging y and z we can obtain a similar inequality for $\hat{z}$. Then

$$\begin{aligned} \frac{\|\hat{z}\|}{\|\hat{y}\|} &\leq \frac{\|z\| + \frac{2r}{1-r}[\|y\| + \|z\|]}{\|y\| - \frac{2r}{1-r}[\|y\| + \|z\|]} \\ &\leq \frac{1+3r}{1-5r}, \end{aligned}$$

since $\|z\| \leq \|y\|$. ∎

REFERENCES

[1] A. W. BOJANCZYK, R. P. BRENT, AND F. DE HOOG, *QR factorization of Toeplitz matrices*, Numer. Math., 49 (1986), pp. 81–94.

[2] 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.

[3] A. W. Bojanczyk, R. P. Brent, P. van Dooren, and F. R. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 210–221.

[4] S. Chandrasekaran and A. H. Sayed, *Stabilizing the generalized Schur algorithm*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 950–983.

[5] J. Chun, *Fast Array Algorithms for Structured Matrices*, Ph.D. thesis, Stanford University, Stanford, CA, 1989.

[6] J. Chun, T. Kailath, and H. Lev-Ari, *Fast parallel algorithms for QR and triangular factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 899–913.

[7] G. Cybenko, *A general orthogonalization technique with applications to time series analysis and signal processing*, Math. Comp., 40 (1983), pp. 323–336.

[8] G. Cybenko, *Fast Toeplitz orthogonalization using inner products*, SIAM J. Sci. Statist. Comp., 8 (1987), pp. 734–740.

[9] I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp., 64 (1995), pp. 1557–1576.

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

[11] M. Gu, *Stable and Efficient Algorithms for Structured Systems of Linear Equations*, Tech. report LBL-37690, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA, 1995.

[12] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, in Linear Algebra for Signal Processing, A. Bojanczyk and G. Cybenko eds., IMA Vol. Math. Appl. 69, Springer, New York, 1995, pp. 63–81.

[13] T. Kailath and J. Chun, *Generalized displacement structure for block-Toeplitz, Toeplitz-block, and Toeplitz-derived matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 114–128.

[14] T. Kailath, S. Y. Kung, and M. Morf, *Displacement ranks of a matrix*, Bull. Amer. Math. Soc., 1 (1979), pp. 769–773.

[15] T. Kailath and A. H. Sayed, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.

[16] A. H. Sayed, *Displacement Structure in Signal Processing and Mathematics*, Ph.D. thesis, Stanford University, Stanford, CA, 1992.

[17] A. H. Sayed and T. Kailath, *A look-ahead block Schur algorithm for Toeplitz-like matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 388–413.

[18] M. Stewart and P. van Dooren, *Stability issues in the factorization of structured matrices*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 104–118.

[19] D. R. Sweet, *Fast Toeplitz orthogonalization*, Numer. Math., 43 (1984), pp. 1–21.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 140–153, January 1998

009

GENERALIZED REFLEXIVE MATRICES: SPECIAL PROPERTIES AND APPLICATIONS*

HSIN-CHU CHEN†

Abstract. The main purpose of this paper is to introduce and exploit special properties of two special classes of rectangular matrices A and B that have the relations

$$A = PAQ \text{ and } B = -PBQ, \qquad A,\ B \in \mathcal{C}^{n\times m},$$

where P and Q are two generalized reflection matrices. The matrices A (B), a generalization of reflexive (antireflexive) matrices and centrosymmetric matrices, are referred to in this paper as generalized reflexive (antireflexive) matrices.

After introducing these two classes of matrices and developing general theories associated with them, we then show how to use some of the important properties to decompose linear least-squares problems whose coefficient matrices are generalized reflexive into two smaller and independent subproblems. Numerical examples are presented to demonstrate their usefulness.

Key words. generalized reflexive matrices, reflexive matrices, centrosymmetric matrices, left orthogonality, right orthogonality

AMS subject classifications. 15A57, 15A90

PII. S0895479895288759

1. Introduction.

In this paper, we introduce two new special classes of matrices A and B that have the following relations:

$$A = PAQ \text{ and } B = -PBQ, \qquad A,\ B \in \mathcal{C}^{n\times m},$$

where P of dimension n and Q of dimension m are two generalized reflection matrices. By a generalized reflection matrix, say R, we mean that R satisfies the following two conditions: (1) $R = R^*$ and (2) $R^2 = I$. In other words, a generalized reflection matrix is an involutory Hermitian matrix. The matrices A and B are, respectively, a generalization of reflexive matrices U and antireflexive matrices V [ChSa87, Chen88], which possess the following special properties:

$$U = PUP \text{ and } V = -PVP, \qquad U,\ V \in \mathcal{C}^{n\times n},$$

where P is some reflection (symmetric signed permutation) matrix. We shall refer to A and B as generalized reflexive and generalized antireflexive matrices, respectively, in this paper. The role of generalized antireflexive matrices to that of generalized reflexive matrices is what antireflexive matrices are to reflexive matrices.

In addition to presenting these two classes of matrices and exploiting their special properties, we provide numerical examples to show how to use the exploited properties to decompose linear least-squares problems with generalized reflexive coefficient matrices into independent subproblems. To illustrate the frequent occurrences of generalized reflexive matrices and the wide applicability of the proposed approach, we also present physical examples obtained from engineering/scientific applications.

* Received by the editors July 7, 1995; accepted for publication (in revised form) by G. P. Styan January 5, 1997. This work was supported by the Louisiana Board of Regents through the Louisiana Education Quality Support Fund LEQSF(1993-95)-RD-A-35. The revision of this paper was supported by the Army Research Laboratory under grant DAAL-03-G-92-0377.

http://www.siam.org/journals/simax/19-1/28875.html

† Army Center of Excellence in Information Sciences, Department of Computer and Information Sciences, Clark Atlanta University, 223 James P. Brawley Dr. at Fair St. SW, Atlanta, GA 30314 (hchen@diamond.cau.edu).

2. Generalized reflexive/antireflexive matrices. In this section, we present some basic definitions related to generalized reflexive matrices. Throughout this paper, we use the superscripts T, $*$, and $+$ to denote the transpose, conjugate transpose, and generalized inverse of matrices (vectors), respectively. All matrix–matrix multiplications and additions are assumed to be conformable if their dimensions are not mentioned.

DEFINITION 2.1. *Let P be some generalized reflection matrix of dimension n.*

- *Reflexive (antireflexive) vectors. A vector $x \in \mathcal{C}^n$ is said to be reflexive (or antireflexive) with respect to P if $x = Px$ (or if $x = -Px$).*
- *Reflexive (antireflexive) matrices. A matrix $A \in \mathcal{C}^{n\times n}$ is said to be reflexive (or antireflexive) with respect to P if $A = PAP$ (or $A = -PAP$).*
- *Reflexive (antireflexive) subspaces. A space S is said to be a reflexive (or antireflexive) subspace with respect to P if every element in S is reflexive (or antireflexive) with respect to the same matrix P.*

It should be mentioned that this definition differs slightly from that in [Chen88, ChSa89a], where reflexivity and antireflexivity are defined in terms of reflection matrices instead of the generalized ones. The main reason for using generalized reflection matrices in this paper to define such special classes of vectors, matrices, and subspaces is that by doing so we not only enlarge their membership but also leave intact all the underlying properties associated with them. Note also that the definition of reflexive subspaces applies both to the spaces consisting of reflexive vectors and to those consisting of reflexive matrices as their elements. The same is true for antireflexive subspaces. When applied to vectors in $\mathcal{C}^n$, reflexive and antireflexive subspaces will be denoted by $\mathcal{C}_r^n(P)$ and $\mathcal{C}_a^n(P)$, respectively. When applied to matrices in $\mathcal{C}^{n\times n}$, they will be denoted by $\mathcal{C}_r^{n\times n}(P)$ and $\mathcal{C}_a^{n\times n}(P)$, respectively.

DEFINITION 2.2. *Let P and Q be two generalized reflection matrices of dimension n and m, respectively.*

- *Generalized reflexive (antireflexive) matrices. A matrix $A \in \mathcal{C}^{n\times m}$ is said to be generalized reflexive (or generalized antireflexive) with respect to the matrix pair (P, Q) if $A = PAQ$ (or $A = -PAQ$).*
- *Generalized reflexive (antireflexive) subspaces. A subspace $S \subset \mathcal{C}^{n\times m}$ is said to be generalized reflexive (or generalized antireflexive) with respect to (P, Q) if $A = PAQ$ (or $A = -PAQ$) for any $A \in S$.*
- *Generalized SAS (anti-SAS) properties. A matrix A is said to possess a generalized SAS (or generalized anti-SAS) property if A is generalized reflexive (or generalized antireflexive), where SAS stands for symmetric and antisymmetric.*

From this definition, it is clear that a reflexive (antireflexive) matrix or vector is necessarily a generalized reflexive (antireflexive) matrix. The converse, however, is not true in general. By taking $P = J_n$ and $Q = J_m$, where J_k is the cross-identity matrix of dimension k, then the generalized reflexive matrices A, $A = PAQ$, reduce to the rectangular centrosymmetric matrices defined in [Weav85]. Therefore, centrosymmetric matrices (square or rectangular), whose special properties have been under extensive study [Zehf62, Aitk49, Good70, Andr73a, Andr73b, PyBA73, CaBu76, Weav88], are also a special case of generalized reflexive matrices.

To serve as an example of generalized reflexive matrices, let

$$A = \begin{bmatrix} \alpha & \beta & \gamma \\ \mu & \nu & \nu \\ \alpha & \gamma & \beta \end{bmatrix}, \; P = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \text{ and } Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Then the matrix A, real or complex, is a generalized reflexive matrix with respect to (P, Q) since $A = PAQ$.

DEFINITION 2.3. *Left (right) orthogonality. A matrix $X \in \mathcal{C}^{n\times m}$ is said to be left (or right) orthogonal to another matrix $Y \in \mathcal{C}^{n\times m}$ if $X^*Y = 0$ (or $YX^* = 0$). Likewise, a subspace S_1 of $\mathcal{C}^{n\times m}$ is said to be left (or right) orthogonal to another subspace S_2 of $\mathcal{C}^{n\times m}$ if $X^*Y = 0$ (or $YX^* = 0$) for any $X \in S_1$ and any $Y \in S_2$.*

It should be pointed out that the conventional orthogonality for vectors is simply the left orthogonality defined above. For simplicity, in the rest of the paper, we shall use L-orthogonality and R-orthogonality to denote left orthogonality and right orthogonality, respectively. Now, let $\mathcal{C}_r^{n\times m}(P, Q)$ and $\mathcal{C}_a^{n\times m}(P, Q)$, where the order of P and Q in the matrix pair is important, be two subsets of the space $\mathcal{C}^{n\times m}$ defined by

$$\mathcal{C}_r^{n\times m}(P, Q) = \{A \mid A \in \mathcal{C}^{n\times m} \text{ and } A = PAQ\}, \tag{1}$$

$$\mathcal{C}_a^{n\times m}(P, Q) = \{A \mid A \in \mathcal{C}^{n\times m} \text{ and } A = -PAQ\}, \tag{2}$$

where P and Q are two generalized reflection matrices of dimension n and m, respectively. Here we use the subscript r (a) to reflect the generalized reflexive (antireflexive) nature of the subsets. Note that if $m = n$ and $Q = P$, then $\mathcal{C}_r^{n\times m}(P, Q)$ and $\mathcal{C}_a^{n\times m}(P, Q)$ reduce to $\mathcal{C}_r^{n\times n}(P)$ and $\mathcal{C}_a^{n\times n}(P)$, respectively. In the case where $m = 1$ and $Q = 1$, $\mathcal{C}_r^{n\times m}(P, Q)$ and $\mathcal{C}_a^{n\times m}(P, Q)$ become $\mathcal{C}_r^{n}(P)$ and $\mathcal{C}_a^{n}(P)$, respectively.

3. Special properties. After presenting the basic definitions for generalized reflexive/antireflexive matrices, we are now in a position to exploit their fundamental properties.

THEOREM 3.1. *Let P and Q be two generalized reflection matrices of dimensions n and m, respectively, and $\alpha,\ \beta \in \mathcal{C}$.*

1. *If A and B are both in $\mathcal{C}_r^{n\times m}(P, Q)$, then*

$$(\alpha A^+ + \beta B^+) \in \mathcal{C}_r^{m\times n}(Q, P),$$

$$(\alpha A^* + \beta B^*) \in \mathcal{C}_r^{m\times n}(Q, P),$$

$$A^*B \in \mathcal{C}_r^{m\times m}(Q), \text{ and } AB^* \in \mathcal{C}_r^{n\times n}(P).$$

2. *If A and B are both in $\mathcal{C}_a^{n\times m}(P, Q)$, then*

$$(\alpha A^+ + \beta B^+) \in \mathcal{C}_a^{m\times n}(Q, P),$$

$$(\alpha A^* + \beta B^*) \in \mathcal{C}_a^{m\times n}(Q, P),$$

$$A^*B \in \mathcal{C}_r^{m\times m}(Q), \text{ and } AB^* \in \mathcal{C}_r^{n\times n}(P).$$

3. *If A is in $\mathcal{C}_r^{n\times m}(P, Q)$ and B is in $\mathcal{C}_a^{n\times m}(P, Q)$, or vice versa, then*

$$(\alpha A^*A + \beta B^*B) \in \mathcal{C}_r^{m\times m}(Q),$$

$$(\alpha AA^* + \beta BB^*) \in \mathcal{C}_r^{n\times n}(P),$$

$$A^*B \in \mathcal{C}_a^{m\times m}(Q), \text{ and } AB^* \in \mathcal{C}_a^{n\times n}(P).$$

Proof. The generalized inverse of a matrix $A \in \mathcal{C}^{n\times m}$ is typically defined to be the unique matrix X that satisfies the following four Moore–Penrose conditions [Moor35, Penr55]:

$$\text{(a) } AXA = A,\ \text{(b) } XAX = X,\ \text{(c) } (AX)^* = AX,\ \text{and (d) } (XA)^* = XA.$$

To prove part 1, we shall first prove that both A^+ and B^+ are in $\mathcal{C}_r^{m\times n}(Q,P)$. Substitution of $A = PAQ$ into the condition (a) with A^+ replacing X yields

$$PAQA^+PAQ = PAQ \tag{3}$$

since A^+ is the generalized inverse of A. Premultiplying and postmultiplying both sides of (3) by P^{-1} and Q^{-1}, respectively, we have

$$AYA = A,$$

where $Y = QA^+P$. Observe that Y satisfies the first Moore–Penrose condition. By using the fact that both P and Q are unitary Hermitian matrices, it can easily be shown that Y also satisfies the other three Moore–Penrose conditions. Therefore, Y is a generalized inverse of A. Since the Moore–Penrose inverse is known to be unique, we conclude that $A^+ = Y$ and, therefore,

$$A^+ = QA^+P \in \mathcal{C}_r^{m\times n}(Q,P).$$

Likewise,

$$B^+ = QB^+P \in \mathcal{C}_r^{m\times n}(Q,P).$$

Accordingly, $(\alpha A^+ + \beta B^+) \in \mathcal{C}_a^{m\times n}(Q,P)$, The proof for the rest requires no further knowledge and is, therefore, omitted. Analogous proof can also be obtained for parts 2 and 3. □

THEOREM 3.2. *Given two generalized reflection matrices P of dimension n and Q of dimension m, any matrix $A \in \mathcal{C}^{n\times m}$ can be decomposed into two parts U and V, $U + V = A$, such that $U \in \mathcal{C}_r^{n\times m}(P,Q)$ and $V \in \mathcal{C}_a^{n\times m}(P,Q)$.*

Proof. Take

$$U = \frac{1}{2}(A + PAQ) \text{ and } V = \frac{1}{2}(A - PAQ) \tag{4}$$

and employ the involutory property $P^2 = I$ and $Q^2 = I$. The proof is trivial and, thus, omitted. □

Two special instances of this theorem can be found in [ChSa89a, ChSa89b], where one is obtained by setting $m = 1$ and $Q = 1$ for vectors and the other is the special case when $m = n$ and $Q = P$ for square matrices.

THEOREM 3.3. *$\mathcal{C}_r^{n\times m}(P,Q)$ is a generalized reflexive subspace and $\mathcal{C}_a^{n\times m}(P,Q)$ is a generalized antireflexive subspace of $\mathcal{C}^{n\times m}$, with respect to (P,Q) over the field $\mathcal{C}$. Furthermore, $\mathcal{C}_r^{n\times m}(P,I)$ is L-orthogonal to $\mathcal{C}_a^{n\times m}(P,I)$ and $\mathcal{C}_r^{n\times m}(I,Q)$ is R-orthogonal to $\mathcal{C}_a^{n\times m}(I,Q)$, where I is the identity matrix of appropriate dimension.*

Proof. (1) $\mathcal{C}_r^{n\times m}(P,Q)$ and $\mathcal{C}_a^{n\times m}(P,Q)$ are subspaces. From Theorem 3.2, it is clear that $\mathcal{C}_r^{n\times m}(P,Q)$ is a nonempty subset of $\mathcal{C}^{n\times m}$. Now let X and Y be two arbitrary elements in $\mathcal{C}_r^{n\times m}(P,Q)$ and $\alpha \in \mathcal{C}$. The matrix $(\alpha X + Y)$ remains in $\mathcal{C}_r^{n\times m}(P,Q)$ since

$$(\alpha X + Y) = P(\alpha X + Y)Q\ .$$

Therefore, $\mathcal{C}_r^{n\times m}(P,Q)$ is a subspace of $\mathcal{C}^{n\times m}$ over the field $\mathcal{C}$. Analogously, $\mathcal{C}_a^{n\times m}(P,Q)$ is also a subspace of $\mathcal{C}^{n\times m}$ over the field $\mathcal{C}$.

(2) Since $\mathcal{C}_r^{n\times m}(P,Q)$ and $\mathcal{C}_a^{n\times m}(P,Q)$ are subspaces, we conclude from (1), (2), and Definition 2.2 that $\mathcal{C}_r^{n\times m}(P,Q)$ is a generalized reflexive subspace and $\mathcal{C}_a^{n\times m}(P,Q)$ is a generalized antireflexive subspace of $\mathcal{C}^{n\times m}$ with respect to (P,Q) over the field $\mathcal{C}$.

(3) $\mathcal{C}_r^{n\times m}(P,I)$ and $\mathcal{C}_a^{n\times m}(P,I)$ are mutually L-orthogonal. For any $X\in\mathcal{C}_r^{n\times m}(P,I)$ and any $Y\in\mathcal{C}_a^{n\times m}(P,I)$ we have

$$X^*Y=(X^*P^*)(-PY)=-X^*Y=0\ ,$$

$$Y^*X=(X^*Y)^*=0\ .$$

Hence, $\mathcal{C}_r^{n\times m}(P,I)$ and $\mathcal{C}_a^{n\times m}(P,I)$ are mutually L-orthogonal. Likewise, $\mathcal{C}_r^{n\times m}(I,Q)$ and $\mathcal{C}_a^{n\times m}(I,Q)$ are mutually R-orthogonal. □

COROLLARY 3.4.

1. *$\mathcal{C}_r^n(P)$ (or $\mathcal{C}_a^n(P)$) is a generalized reflexive (or generalized antireflexive) subspace of $\mathcal{C}^n$ with respect to P over the field $\mathcal{C}$. Furthermore, $\mathcal{C}_r^n(P)$ and $\mathcal{C}_a^n(P)$ are mutually L-orthogonal.*
2. *$\mathcal{C}_r^{n\times n}(P)$ (or $\mathcal{C}_a^{n\times n}(P)$) is a generalized reflexive (or generalized antireflexive) subspace of $\mathcal{C}^{n\times n}$ with respect to P over the field $\mathcal{C}$.*

This is trivial because a reflexive (antireflexive) subspace is necessarily a generalized reflexive (generalized antireflexive) subspace. Note that $\mathcal{C}_r^n(P)$ and $\mathcal{C}_a^n(P)$ are not mutually R-orthogonal in general although they are mutually L-orthogonal.

THEOREM 3.5. *Given a linear least-squares problem*

$$\min_x \|Ax-b\|_2,\ A\in\mathcal{C}^{n\times m},\ x\in\mathcal{C}^m,\ b\in\mathcal{C}^n,\ m\le n,$$

where A is assumed to have full column rank, i.e., $rank(A)=m$, let $\tilde{x}$ be the unique solution to the problem and $\tilde{r}=b-A\tilde{x}$, the residual.

1. *If $A\in\mathcal{C}_r^{n\times m}(P,Q)$, then*

$$\tilde{x}\in\mathcal{C}_r^m(Q)\ and\ \tilde{r}\in\mathcal{C}_r^n(P)\ if\ b\in\mathcal{C}_r^n(P), \tag{5}$$

$$\tilde{x}\in\mathcal{C}_a^m(Q)\ and\ \tilde{r}\in\mathcal{C}_a^n(P)\ if\ b\in\mathcal{C}_a^n(P). \tag{6}$$

2. *If $A\in\mathcal{C}_a^{n\times m}(P,Q)$, then*

$$\tilde{x}\in\mathcal{C}_r^m(Q)\ and\ \tilde{r}\in\mathcal{C}_a^n(P)\ if\ b\in\mathcal{C}_a^n(P), \tag{7}$$

$$\tilde{x}\in\mathcal{C}_a^m(Q)\ and\ \tilde{r}\in\mathcal{C}_r^n(P)\ if\ b\in\mathcal{C}_r^n(P). \tag{8}$$

Proof. The proof for part 2 is analogous to that for part 1. Therefore, we need only prove part 1. From the assumption that $A\in\mathcal{C}_r^{n\times m}(P,Q)$, we have $A=PAQ$, where P and Q are, by definition, generalized reflection matrices and thus

$$P=P^*=P^{-1}\text{ and }Q=Q^*=Q^{-1}.$$

Since $\text{rank}(A)=m$, A^+ can be expressed as

$$A^+=(A^*A)^{-1}A^*=(QA^*PPAQ)^{-1}QA^*P$$

$$= (QA^*AQ)^{-1}QA^*P = Q(A^*A)^{-1}QQA^*P$$

$$= QA^+P.$$

It follows that if $b = Pb$, we have

$$\tilde{x} = A^+b = QA^+Pb = QA^+b = Q\tilde{x}$$

and

$$\tilde{r} = b - A\tilde{x} = Pb - PAQQ\tilde{x} = P(b - A\tilde{x}) = P\tilde{r}.$$

Analogously, if $b = -Pb$, then

$$\tilde{x} = A^+b = QA^+Pb = -QA^+b = -Q\tilde{x}$$

and

$$\tilde{r} = b - A\tilde{x} = -Pb - PAQ(-Q\tilde{x}) = -P(b - A\tilde{x}) = -P\tilde{r}.$$

This completes our proof. □

Remark 1. Note that the converse of (5), (6), (7), and (8) does not hold in general. For example, let b be some vector that is neither reflexive nor antireflexive with respect to P, i.e., $b \notin \mathcal{C}_r^n(P)$ and $b \notin \mathcal{C}_a^n(P)$. From a special application of Theorem 3.2 for vectors, however, we can decompose b into b_1 and b_2, both nonzero, such that $b_1 \in \mathcal{C}_r^n(P)$ and $b_2 \in \mathcal{C}_a^n(P)$. Let

$$\tilde{x}_1 = A^+b_1 \text{ and } \tilde{x}_2 = A^+b_2.$$

We have

$$\tilde{x} = A^+b = A^+(b_1 + b_2) = \tilde{x}_1 + \tilde{x}_2,$$

where $\tilde{x}_1 \in \mathcal{C}_r^m(Q)$ and $\tilde{x}_2 \in \mathcal{C}_a^m(Q)$. Now, if b_2 lies in the null space of A^*, then

$$\tilde{x} = \tilde{x}_1 \in \mathcal{C}_r^m(Q)$$

since $A^+b_2 = (A^*A)^{-1}A^*b_2 = 0$. Likewise, if b_1 lies in the null space of A^*, then

$$\tilde{x} = \tilde{x}_2 \in \mathcal{C}_a^m(Q).$$

Remark 2. If the matrix A is rank deficient, then least-squares solutions other than the minimal-norm solution might not have the property shown in this theorem. It is not difficult to construct such examples. One of the simplest cases would be to allow A to have some columns filled with zeros.

Theorems 3.2 and 3.5 yield all the important information we need to decompose a linear least-square problem whose coefficient matrix is generalized reflexive with respect to (P, Q) into two independent and smaller subproblems once P and Q are known. Two simple examples are given in the next section to demonstrate the basic ideas.

To close this section, we show that the generalized SAS (or anti-SAS) property of a matrix is invariant under any signed permutations. By signed permutation we mean that in addition to the permutation of given rows (columns) of the matrix, the sign of

the elements in the rows (columns) could also be reversed. Now, suppose that a matrix A lies in $\mathcal{C}_r^{n\times m}(P,Q)$ so that A is generalized reflexive with respect to to (P,Q). Let M_P and M_Q be two signed permutation matrices whose dimensions are the same as those of P and Q, respectively. Note that M_P and M_Q are not necessarily reflection matrices. They are, however, orthogonal. Denoting $M_P^TAM_Q$ by $\tilde{A}$, $M_P^TPM_P$ by $\tilde{P}$, and $M_Q^TQM_Q$ by $\tilde{Q}$, we have

$$\tilde{A} = M_P^TAM_Q = M_P^TPAQM_Q = M_P^TP(M_PM_P^T)A(M_QM_Q^T)QM_Q$$

$$= (M_P^TPM_P)(M_P^TAM_Q)(M_Q^TQM_Q)$$

$$= \tilde{P}\tilde{A}\tilde{Q},$$

where we have employed the orthogonality of M_P and M_Q. The matrices $\tilde{P}$ and $\tilde{Q}$, like P and Q, are again generalized reflection matrices since

$$\tilde{P} = \tilde{P}^* = \tilde{P}^{-1} \text{ and } \tilde{Q} = \tilde{Q}^* = \tilde{Q}^{-1}.$$

Therefore, the resulting matrix $\tilde{A}$ remains generalized reflexive. In fact, the above argument can be generalized to include unitary transformations, i.e., if A is generalized reflexive with respect to (P,Q), then the matrix $\hat{A}$, $\hat{A} = U_P^*AU_Q$ with U_P and U_Q being unitary, is generalized reflexive with respect to $(\hat{P},\hat{Q})$, where $\hat{P} = U_P^*PU_P$ and $\hat{Q} = U_Q^*QU_Q$.

4. Numerical examples. In this section, we provide examples to demonstrate how to take advantage of the special properties developed in the previous section for handling a special class of linear least-squares problems. The first example presented is a reflexive overdetermined linear system where not only is the matrix A generalized reflexive but the vector b is reflexive as well.

Example 1. Consider the linear least-squares solution to the following overdetermined linear system:

$$Ax = b, \quad \text{where } A = \begin{bmatrix} 4 & 2 \\ 1 & 3 \\ 2 & 4 \\ 3 & 1 \end{bmatrix}, \; x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \text{ and } b = \begin{bmatrix} 16 \\ 15 \\ 16 \\ 15 \end{bmatrix}. \tag{9}$$

Let

$$P = \begin{bmatrix} 0 & I_2 \\ I_2 & 0 \end{bmatrix} \text{ and } Q = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

where I_2 is the identity matrix of dimension 2. It is easy to see that $A = PAQ$ and $b = Pb$. From Theorem 3.5 we know that $x = Qx$, i.e., $x_1 = x_2$. Solving (9) is, therefore, equivalent to solving

$$\begin{bmatrix} 6 \\ 4 \end{bmatrix} \begin{bmatrix} x_1 \end{bmatrix} = \begin{bmatrix} 16 \\ 15 \end{bmatrix}. \tag{10}$$

The least-squares solution to (10) is $x_1 = 3$. Accordingly, the solution to the original problem is $x_1 = x_2 = 3$, which can be verified by solving the normal equation $A^TAx =$

$A^T b$. The residual r, $r = b - Ax = [-2,\ 3,\ -2,\ 3\ ]^T$, is obviously reflexive with respect to P, as expected, since $r = Pr$.

It deserves mentioning that the generalized reflexivity property of the matrix A usually comes from physical models with some sort of reflexive symmetry. The vector b, nevertheless, could be arbitrary and will not have any special form in general. This, however, should not impose any difficulty since given P, any vector can be decomposed into a reflexive and an antireflexive part. Once the decomposition is performed, Theorem 3.5 can be employed to take advantage of the reflexivity and antireflexivity present in the problem, as shown in the next example where we choose b to be neither reflexive nor antireflexive.

Example 2. In this example, we solve the same problem as shown in Example 1, except that the right-hand-side vector b is now taken to be $b = [19,\ 14,\ 13,\ 16\ ]^T$, which is neither reflexive nor antireflexive. To use the generalized reflexivity property of A, we first decompose b into u and v so that $u = Pu$ (reflexive) and $v = -Pv$ (antireflexive). The decomposition yields $u = [16,\ 15,\ 16,\ 15]^T$ and $v = [3,\ -1,\ -3,\ 1]^T$. Instead of solving $Ax = b$ directly, one can always solve $Ay = u$ for y and $Az = v$ for z to obtain x since $x = A^+ b = A^+(u + v) = A^+ u + A^+ v = y + z$. Without exploiting the generalized SAS property of A, however, this decomposition would not offer any advantage since it doubles the amount of computational work.

Our next step is to use Theorem 3.5 to reduce the size of both $Ay = u$ and $Az = v$. Since $Ay = u$ in this example is exactly the same as $Ax = b$ in Example 1, no further demonstration is necessary in this part; the solution is simply $y = [\ 3,\ 3]^T$. The decomposition for $Az = v$ is similar to that for $Ay = u$ except now we have to use the antireflexivity of v. From Theorem 3.5, we know that $z = -Pz$ since $A = PAQ$ and $v = -Qv$. Letting $z = [z_1,\ z_2]^T$, we have $z_1 = -z_2$. Therefore, solving $Az = v$ reduces to solving

$$\begin{bmatrix} 2 \\ -2 \end{bmatrix} \begin{bmatrix} z_1 \end{bmatrix} = \begin{bmatrix} -3 \\ 1 \end{bmatrix}$$

whose least-squares solution is $z_2 = -1$, yielding $z = [\ 1,\ -1]^T$ since $z_1 = -z_2$. Therefore, the solution to the original problem is

$$x = y + z = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

which can again be verified by solving the normal equation of the original system. At this point, it is clear that by employing the special generalized SAS property of A we can decompose the system into two smaller and independent subsystems to solve. This completes our demonstration.

5. Applications to physical problems. As mentioned earlier, generalized reflexive matrices, including the special case of reflexive matrices, usually arise from physical problems with some form of reflexive symmetry. The examples provided in the previous section are chosen arbitrarily for demonstration purposes, without any association with engineering/scientific applications. In this section, we present three examples that are obtained from physical problems in three different application areas; one deals with the altitude estimation of a level network which yields a linear least-squares problem, the second is an electric network resulting in a linear system, and the third problem arises from structural analysis of trusses. In addition to giving the matrices, we shall also briefly describe the physical problems that give rise to such

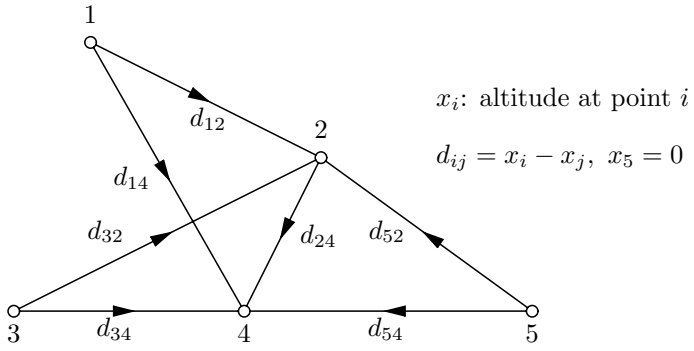


FIG. 1. *A level network for altitude estimation.*

matrices and solve the level network problem. The examples to be given are small. However, they clearly demonstrate the frequent occurrences of generalized reflexive matrices and illustrate the benefit and wide applicability of the proposed approach.

Example 3. In this example, we solve a linear least-squares problem that results from the level network shown in Figure 1, on which the difference in altitude is measured. The objective is to estimate the height above sea level for points 1, 2, 3, and 4. Point 5 is known to lie at sea level.

Let $d_{ij} = x_i - x_j$, where x_i and x_j are the heights above sea level for points i and j, respectively, $i,\ j = 1, \ldots, 5$. This problem yields the following overdetermined linear system:

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} d_{12} \\ d_{52} \\ d_{14} \\ d_{24} \\ d_{34} \\ d_{54} \\ d_{32} \end{bmatrix}. \tag{11}$$

Denoting this system by $Ax = b$, with b assumed to take the numerical values

$$b = \begin{bmatrix} 50 & -152 & 78 & 33 & 30 & -123 & 2 \end{bmatrix}^T,$$

we first observe that the coefficient matrix A is a generalized reflexive matrix: $A = PAQ$ with

$$P = \begin{bmatrix} 0 & 0 & I_3 \\ 0 & -1 & 0 \\ I_3 & 0 & 0 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} 0 & I_2 \\ I_2 & 0 \end{bmatrix}. \tag{12}$$

Note that the coefficient matrix A of this overdetermined system is simply the edge-node incidence matrix of the network.

There are two approaches to solving overdetermined linear systems, one forming its normal equation directly and the other using a QR decomposition instead, mainly for stability reasons. In either approach, we decompose the problem into two independent subproblems first by taking advantage of the generalized reflexivity property of A. Decomposing b into u and v such that $u = Pu$ and $v = -Pv$ and then solving

$Ay = u$ and $Az = v$ for y and z using the reduced systems, as was done in Example 2, yield

$$(13)\quad \begin{bmatrix} 1 & -1 \\ 0 & -1 \\ 1 & -1 \\ 0 & 0 \\ 1 & -1 \\ 0 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 40.0 \\ -137.5 \\ 40.0 \\ 0.0 \\ 40.0 \\ -137.5 \\ 40.0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & -1 \\ 0 & -1 \\ 1 & 1 \\ 0 & 2 \\ -1 & 1 \\ 0 & 1 \\ -1 & -1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 10.0 \\ -14.5 \\ 38.0 \\ 33.0 \\ -10.0 \\ 14.5 \\ -38.0 \end{bmatrix}$$

which are now equivalent to

$$(14)\quad \begin{bmatrix} 1 & -1 \\ 0 & -1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 40.0 \\ -137.5 \\ 40.0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & -1 \\ 0 & -1 \\ 1 & 1 \\ 0 & \frac{2}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 10.0 \\ -14.5 \\ 38.0 \\ \frac{33.0}{\sqrt{2}} \end{bmatrix},$$

respectively, since the last three rows of (13) are identical to the first three in both systems. Note that in removing these equations, the fourth one in the second reduced system must be scaled by a factor of $\sqrt{2}$ as shown in (14) so that the 2-norm of the right-hand side in (14) is consistent with that in (13), i.e.,

$$\| v \|_2^2 = 2\left(v_1^2 + v_2^2 + v_3^2 + \left(\frac{v_4}{\sqrt{2}}\right)^2\right).$$

This scaling is necessary and in fact comes from premultiplying both sides of the system by the orthogonal matrix X:

$$X = \frac{1}{\sqrt{2}} \begin{bmatrix} I_3 & 0 & -I_3 \\ 0 & \sqrt{2} & 0 \\ I_3 & 0 & I_3 \end{bmatrix},$$

where I_3 is an identity matrix of order 3. The scaling has no effect on the fourth equation in the first reduced system since it is a dummy and can simply be removed.

The solution to (11) can now be obtained with ease from (14) which gives

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 177.5 \\ 137.5 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 24.0 \\ 15.1 \end{bmatrix},$$

yielding

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ -z_1 \\ -z_2 \end{bmatrix} = \begin{bmatrix} 201.5 \\ 152.6 \\ 153.5 \\ 122.4 \end{bmatrix}.$$

The correctness of this solution can be verified by solving the normal equation of the original system, which obviously requires many more arithmetic operations than this approach.

In the following two examples, we use J_k to denote the cross-identity matrix of dimension k:

$$J_k(i,j) = \begin{cases} 1 & \text{if } i + j = k + 1,\ i,\ j = 1, \ldots, k, \\ 0 & \text{otherwise.} \end{cases}$$

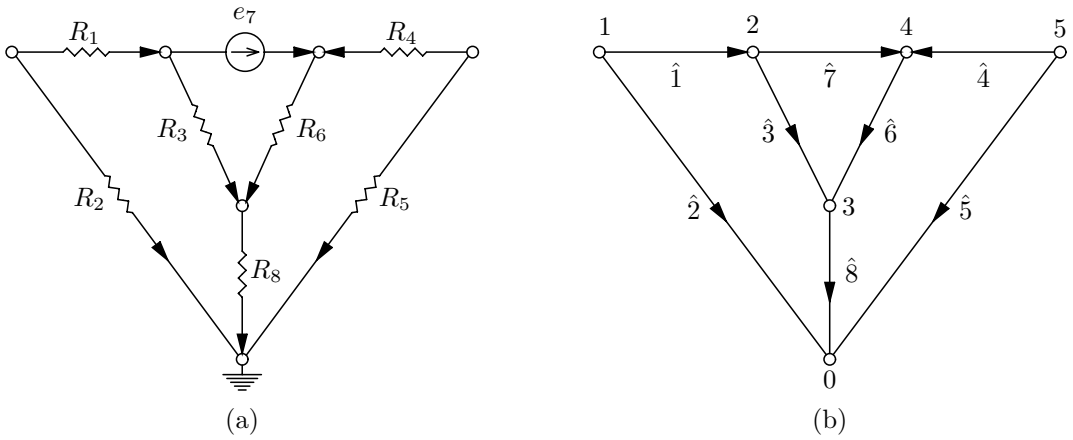


FIG. 2. *An electrical network* (a) *and its graph numbering* (b).

The identity matrix of dimension k is denoted by I_k, as usual.

Example 4. We now consider the electrical network shown in Figure 2, where R_i, the resistances of resistors, and e_7, a current source, are constants. The matrix A given below is the node-edge incidence matrix (or the transpose of the edge-node incidence matrix) associated with this network:

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

This incidence matrix is involved in the well-known Kirchhoff's current law (KCL) and Kirchhoff's voltage law (KVL) equations for circuits: $Ae_b = 0$ and $A^T v_n = v_b$, respectively, where e_b is the branch current vector, v_b is the branch voltage vector, and v_n is the node voltage vector [Dire75, pp. 283–291].

Let P and Q be the following two reflection matrices:

$$P = J_5 \quad \text{and} \quad Q = \begin{bmatrix} 0 & I_3 & & \\ I_3 & 0 & & \\ & & -1 & 0 \\ & & 0 & 1 \end{bmatrix},$$

where the unshown entries in the matrix are zero. It is simple practice to observe that $A = PAQ$, i.e., generalized reflexive with respect to (P, Q). Let Y be the diagonal matrix that represents the physical properties of the circuit. In our example,

$$Y = \text{diag}\left(\frac{1}{R_1}, \frac{1}{R_2}, \frac{1}{R_3}, \frac{1}{R_4}, \frac{1}{R_5}, \frac{1}{R_6}, 0, \frac{1}{R_8}\right).$$

The branch current-voltage relationship can be expressed as

$$e_b = Y v_b + \varphi_b, \tag{15}$$

where φ_b is a vector representing branch current sources. Substitution of v_b in the KVL equation into (15) and then e_b into the KCL equation yields the following linear

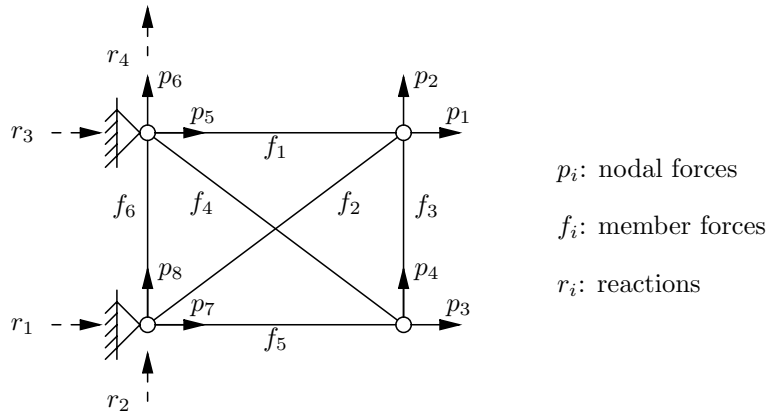


FIG. 3. *Force components of a four-node six-member rectangular truss structure.*

system:

$$AYA^T v_n = b_n, \quad \text{where} \quad b_n = -A\varphi_b. \tag{16}$$

If $R_1 = R_4$, $R_2 = R_5$, and $R_3 = R_6$, then the matrix Y is reflexive with respect to Q since $Y = QYQ$. This will be the case when this circuit has reflexive symmetry. Accordingly, the system admittance matrix $K = AYA^T$ is reflexive with respect to P. In this case, K is also a centrosymmetric matrix. Whatever b_n is, (16) can be decomposed into two decoupled subsystems using the same approach already discussed. Details are omitted.

It deserves mentioning that the reflection matrix P is obtained solely from the numbering of nodes, yet Q is determined by both the numbering and the orientation of the branches of the graph shown in Figure 2. Reordering of nodes and branches or reorientation of branches will certainly change the incidence matrix A. The (generalized) SAS property associated with the problem, however, will never be destroyed. All we need to change are just the reflection matrices P and Q.

Example 5. In our last example, we give matrices that arise in the stress analysis of a rectangular truss structure employing the force method. The truss and its force components are shown in Figure 3, where the horizontal members are assumed to be four feet long and the vertical members are three feet long.

Let f, p, and r be the vectors that consist of the element forces f_i, external nodal force components p_i, and reaction components r_i, respectively. The force equilibrium equations of a structure can generally be expressed as [Prze68, pp. 206–209]

$$\begin{bmatrix} A_f & A_r \end{bmatrix} \begin{bmatrix} f \\ r \end{bmatrix} = p,$$

where A_f and A_r are rectangular matrices whose coefficients are the direction cosines relating the element forces f and the reaction components r, respectively, to the

external forces p. In our example, this relation gives rise to the following matrix A:

$$A = \left[\begin{array}{c|c} A_f & A_r \end{array}\right] = \left[\begin{array}{cccccc|cccc} 1 & c & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & s & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & -s & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & -c & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & s & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & -c & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -s & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \end{array}\right], \quad \begin{array}{l} c = \frac{4}{5}, \\ s = \frac{3}{5}. \end{array}$$

Let P, Q_f, and Q_r be the reflection matrices given below:

$$P = \begin{bmatrix} 0 & S_2 & & \\ S_2 & 0 & & \\ & & 0 & S_2 \\ & & S_2 & 0 \end{bmatrix}, \; Q_f = \begin{bmatrix} J_5 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } \; Q_r = \begin{bmatrix} 0 & S_2 \\ S_2 & 0 \end{bmatrix},$$

where $S_2 = \text{diag } (1, \; -1)$. We have

$$A_f = PA_fQ_f, \; A_r = PA_rQ_r, \text{ and } \; A = PAQ,$$

where

$$Q = \begin{bmatrix} Q_f & 0 \\ 0 & Q_r \end{bmatrix}.$$

In other words, the matrices A_f, A_r, and A are all generalized reflexive matrices and, therefore, the proposed approach applies.

6. Conclusions. In this paper, we have introduced two new special classes of rectangular matrices A and B, $A, \; B \in \mathcal{C}^{n\times m}$, that have the relations

$$A = PAQ \text{ and } B = -PBQ,$$

where P and Q are two generalized reflection matrices. They are generalizations of reflexive matrices U and antireflexive matrices V, $U, \; V \in \mathcal{C}^{n\times n}$, that take the form

$$U = PUP \text{ and } V = -PVP$$

and, therefore, have more general properties. The matrices A (B) are referred to as generalized reflexive (antireflexive) matrices in this paper. Many computationally important and interesting special properties associated with these new classes of matrices have been exploited. This exploitation allows linear least-squares problems (and linear systems as well) with generalized reflexive coefficient matrices to be decomposed into smaller and independent subproblems to solve, yielding computational efficiency and large-grain parallelism at the same time. An efficient decomposition method based on the exploited special properties has also been proposed and illustrated.

Although it is not trivial to realize the existence of generalized reflexive matrices from the matrix point of view, this new class of matrices indeed arise very often in many application disciplines. In this paper, three examples obtained from physical problems in distinct application areas have been presented to demonstrate their frequent occurrences. Our investigation indicates that generalized reflexive matrices arise naturally from problems with reflexive symmetry, which account for a great number of real-world scientific and engineering applications. Therefore, the proposed efficient approach will certainly have widely important applications.

REFERENCES

[Aitk49] A. C. Aitken, *Determinants and Matrices*, 6th ed., Wiley-Interscience, New York, 1949.

[Andr73a] A. L. Andrew, *Solution of equations involving centrosymmetric matrices*, Technometrics, 15 (1973), pp. 405–407.

[Andr73b] A. L. Andrew, *Eigenvectors of certain matrices*, Linear Algebra Appl., 7 (1973), pp. 151–162.

[CaBu76] A. Cantoni and P. Butler, *Eigenvalues and eigenvectors of symmetric centrosymmetric matrices*, Linear Algebra Appl., 13 (1976), pp. 275–288.

[Chen88] H.-C. Chen, *The SAS Domain Decomposition Method for Structural Analysis*, CSRD Tech. report 754, Center for Supercomputing Research and Development, University of Illinois, Urbana, IL, 1988.

[ChSa87] H.-C. Chen and A. Sameh, *Numerical linear algebra algorithms on the cedar system*, in Parallel Computations and Their Impact on Mechanics, A. K. Noor, ed., The American Society of Mechanical Engineers, AMD-Vol. 86, 1987, pp. 101–125.

[ChSa89a] H.-C. Chen and A. Sameh, *A matrix decomposition method for orthotropic elasticity problems*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 39–64.

[ChSa89b] H.-C. Chen and A. Sameh, *A domain decomposition method for 3D elasticity problems*, in Applications of Supercomputers in Engineering: Fluid Flow and Stress Analysis Applications, C. A. Brebbia and A. Peters, eds., Computational Mechanics Publications, Southampton University, Southampton, UK, 1989, pp. 171–188.

[Dire75] S. W. Director, *Circuit Theory: A Computational Approach*, John Wiley, New York, 1975.

[Good70] I. J. Good, *The inverse of a centrosymmetric matrix*, Technometrics, 12 (1970), pp. 925–928.

[Moor35] E. H. Moore, *General Analysis*, Mem. Amer. Philos. Soc., I, 1935, pp. 147–209.

[Penr55] R. Penrose, *A generalized inverse of matrices*, Proc. Cambridge Philos. Soc., 51 (1955), pp. 406–413.

[Prze68] J. S. Przemieniecki, *Theory of Matrix Structural Analysis*, McGraw–Hill, New York, 1968.

[PyBA73] W. C. Pye, T. L. Boullino, and T. A. Atchison, *The pseudoinverse of a centrosymmetric matrix*, Linear Algebra Appl., 6 (1973), pp. 201–204.

[Weav85] J. R. Weaver, *Centrosymmetric (cross-symmetric) matrices, their basic properties, eigenvalues, and eigenvectors*, Amer. Math. Monthly, 92 (1985), pp. 711–717.

[Weav88] J. R. Weaver, *Real eigenvalues of nonnegative matrices which commute with a symmetric matrix involution*, Linear Algebra Appl., 110 (1988), pp. 243–253.

[Zehf62] G. Zehfuss, *Zwei Sätze über determinanten*, Zeitschrift f. Math. u. Phys., vii. (1862), pp. 436–439.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 154–160, January 1998

010

EFFICIENT SOLUTION OF CONSTRAINED LEAST SQUARES PROBLEMS WITH KRONECKER PRODUCT STRUCTURE*

ANDERS BARRLUND†

Abstract. A computational method for efficient solution of linear constrained least squares problems with Kronecker product structure is presented. The equality constraints are assumed to be linearly independent. The computational efficiency of the method is analyzed. Conditions for uniqueness of solutions are given.

Key words. Kronecker product, constrained least squares, generalized singular value decomposition

AMS subject classifications. 65F20, 65F30, 15A57, 65K10

PII. S0895479895295027

1. Introduction.

In this paper we consider the constrained least squares problem

$$\min_x \| (A_1 \otimes A_2)x - f\|_2 \tag{1.1}$$

$$\text{subject to } (B_1 \otimes B_2)x = g,$$

where $A_1 \in R^{m_{a1}\times n_{a1}}$, $A_2 \in R^{m_{a2}\times n_{a2}}$, $B_1 \in R^{m_{b1}\times n_{b1}}$, $B_2 \in R^{m_{b2}\times n_{b2}}$, $n_{a1} = n_{b1}$, $n_{a2} = n_{b2}$, $f \in R^{m_{a1}m_{a2}}$, the rows of $B_1 \otimes B_2$ are linearly independent, and the solution $x \in R^{n_{a1}n_{a2}}$. Note that $B_1 \otimes B_2$ has independent rows if and only if B_1 and B_2 have linearly independent rows. One example of a practical application where these problems occur is in surface fitting with hierarchical splines [3, section 5]. It is possible to take advantage of the Kronecker product structure to get an efficient algorithm for the solution of the problem (1.1). The solution of the problem (1.1) is not always unique. There exist cases where we can guarantee that the solution is not unique only by looking at the dimensions of A_1, A_2, B_1, and B_2. These cases include some examples where $m_{a1}m_{a2} + m_{b1}m_{b2} > n_{b1}n_{b2}$.

In [2] and [7] it is studied how unconstrained linear least squares problems, involving Kronecker products, can be solved efficiently. Forsey and Bartels [3] present a way to transform (1.1) to the form

$$A_2V_ZA_1^T \approx F - A_2V_WA_1^T \tag{1.2}$$

(formula (40) in [3]), where V_W is fixed by the constraints and V_Z satisfies $B_2V_ZB_1^T = 0$. However, any method that takes advantage of Kronecker product structure, to solve the problem (1.2) subject to $B_1V_ZB_1^T = 0$, cannot be found in [3]. The new results in this paper show how to take advantage of both the Kronecker product in the constraints and the Kronecker product in the least squares equations when solving (1.1).

The paper is outlined as follows: section 2 describes the computational method to solve constrained least squares problems with Kronecker product structure. Section

* Received by the editors November 22, 1995; accepted for publication (in revised form) by G. Cybenko January 6, 1997.
http://www.siam.org/journals/simax/19-1/29502.html
† Department of Computing Science, Umeå University, S-90187 Umeå, Sweden (abrr@cs.umu.se).

3 examines in which cases the solution is unique and not unique. In section 4, it is shown how the generalized singular value decomposition (GSVD) can be applied to these problems. In section 5, the complexity of the method is analyzed. Finally, in section 6 we make some conclusions. We will use the following notation: $\otimes$ denotes the Kronecker product; $\|\cdot\|_2$ denotes the Euclidean vector norm; $\|\cdot\|_F$ denotes the Frobenius norm; the superscripts T and -1 denote the transpose and the inverse, respectively; $x(i{:}j)$ denotes elements i to j of vector x; $x(i{:}j{:}k)$ denotes elements $i, i+j, i+2j, ..., k$ of vector x; $A(:,i{:}j)$ denotes columns i to j of matrix A. $A(i{:}j,k{:}l)$ denotes a submatrix consisting of rows i to j and columns k to l of matrix A (i.e., MATLAB style); I_n denotes the $n \times n$ identity matrix. $vec(A)$ is the elements of A placed in one column vector, as $vec(A) = [A(:,1)^T A(:,2)^T \ldots]^T$.

2. The method to solve linear constrained least squares problems with Kronecker product structure. It is well known [3], [4, p. 25] that the equation

$$(A \otimes B)vec(X) = vec(F)$$

is equivalent to the equation

$$BXA^T = F.$$

This equivalence implies that (1.1) can be rewritten as

$$\min_X \| A_2 X A_1^T - F\|_F \tag{2.1}$$

$$\text{subject to } B_2 X B_1^T = G,$$

where $vec(X) = x$, $vec(F) = f$, and $vec(G) = g$. The approach we suggest is a null space method [1, pp. 190–191], [6, Chapter 20]. We perform a change of variables that makes it possible to divide the unknowns into two sets. One is the set determined by the constraints and the other is the set belonging to the null space of the constraints.

Let $[L_{b1}\ 0]Q_{b1} = B_1$ and $[L_{b2}\ 0]Q_{b2} = B_2$ be the LQ factorizations of B_1 and B_2 (given by the QR factorizations of B_1^T and B_2^T). Let $\tilde{A}_1 = A_1 Q_{b1}^T$ and $\tilde{A}_2 = A_2 Q_{b2}^T$. With the change of variables

$$\begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} = Y = Q_{b2} X Q_{b1}^T,$$

where $Y_{11} \in R^{m_{b2} \times m_{b1}}$, $Y_{12} \in R^{m_{b2} \times (n_{b1} - m_{b1})}$, $Y_{21} \in R^{(n_{b2} - m_{b2}) \times m_{b1}}$, and $Y_{22} \in R^{(n_{b2} - m_{b2}) \times (n_{b1} - m_{b1})}$, the problem (2.1) takes the form

$$\min_Y \| \tilde{A}_2 Y \tilde{A}_1^T - F\|_F \tag{2.2}$$

$$\text{subject to } L_{b2} Y_{11} L_{b1}^T = G.$$

The submatrix Y_{11} is determined by the constraints. In the remaining steps, we can consider Y_{11} as a known matrix.

The rest of the unknowns, Y_{12}, Y_{21}, and Y_{22} belong to the null space of the constraints. Let $\hat{A}_1 = [\tilde{A}_1(:, m_{b1}+1 : n_{b1})\ \ \tilde{A}_1(:, 1 : m_{b1})]$ and $\hat{A}_2 = [\tilde{A}_2(:, m_{b2}+1 : n_{b2})\ \ \tilde{A}_2(:, 1 : m_{b2})]$. Let

$$\hat{A}_1 = Q_{a1} R^{(a1)}, \qquad R^{(a1)} = \begin{bmatrix} R_{11}^{(a1)} & R_{12}^{(a1)} \\ 0 & R_{22}^{(a1)} \end{bmatrix}, \tag{2.3}$$

$$\hat{A}_2 = Q_{a2} R^{(a2)}, \qquad R^{(a2)} = \begin{bmatrix} R_{11}^{(a2)} & R_{12}^{(a2)} \\ 0 & R_{22}^{(a2)} \end{bmatrix}$$

be the "economy size" QR factorizations of $\hat{A}_1$ and $\hat{A}_2$. (In the economy size QR factorization $Q \in R^{m\times r}$ and $R \in R^{r\times n}$, where $r = \min(m,n)$). In (2.3), $R_{11}^{(a1)} \in R^{q_{a1}\times(n_{b1}-m_{b1})}$, $R_{12}^{(a1)} \in R^{q_{a1}\times m_{b1}}$, $R_{22}^{(a1)} \in R^{(r_{a1}-q_{a1})\times m_{b1}}$, $R_{11}^{(a2)} \in R^{q_{a2}\times(n_{b2}-m_{b2})}$, $R_{12}^{(a2)} \in R^{q_{a2}\times m_{b2}}$, $R_{22}^{(a2)} \in R^{(r_{a2}-q_{a2})\times m_{b2}}$, $Q_{a1} \in R^{m_{a1}\times r_{a1}}$, and $Q_{a2} \in R^{m_{a2}\times r_{a2}}$, where $r_{a1} = \min(m_{a1}, n_{a1})$, $r_{a2} = \min(m_{a2}, n_{a2})$, $q_{a1} = \min(r_{a1}, n_{b1} - m_{b1})$, and $q_{a2} = \min(r_{a2}, n_{b2} - m_{b2})$. Then (2.2) has the same solution as

$$\min_{Y_{12},Y_{21},Y_{22}} \left\| \begin{bmatrix} R_{11}^{(a2)} & R_{12}^{(a2)} \\ 0 & R_{22}^{(a2)} \end{bmatrix} \begin{bmatrix} Y_{22} & Y_{21} \\ Y_{12} & Y_{11} \end{bmatrix} \begin{bmatrix} R_{11}^{(a1)T} & 0 \\ R_{12}^{(a1)T} & R_{22}^{(a1)T} \end{bmatrix} - \tilde{F} \right\|_F, \tag{2.4}$$

where

$$\tilde{F} = \begin{bmatrix} \tilde{F}_{11} & \tilde{F}_{12} \\ \tilde{F}_{21} & \tilde{F}_{22} \end{bmatrix} = Q_{a2}^T F Q_{a1},$$

where $\tilde{F}_{11} \in R^{q_{a2}\times q_{a1}}$, $\tilde{F}_{12} \in R^{q_{a2}\times(r_{a1}-q_{a1})}$, $\tilde{F}_{21} \in R^{(r_{a2}-q_{a2})\times q_{a1}}$, and $\tilde{F}_{22} \in R^{(r_{a2}-q_{a2})\times(r_{a1}-q_{a1})}$. Block multiplication in (2.4) gives

$$\min \left\| \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \right\|_F,$$

where

$$M_{11} = R_{11}^{(a2)} Y_{22} R_{11}^{(a1)T} + R_{12}^{(a2)} Y_{12} R_{11}^{(a1)T} + R_{11}^{(a2)} Y_{21} R_{12}^{(a1)T} + R_{12}^{(a2)} Y_{11} R_{12}^{(a1)T} - \tilde{F}_{11},$$

$$M_{12} = R_{11}^{(a2)} Y_{21} R_{22}^{(a1)T} + R_{12}^{(a2)} Y_{11} R_{22}^{(a1)T} - \tilde{F}_{12},$$

$$M_{21} = R_{22}^{(a2)} Y_{12} R_{11}^{(a1)T} + R_{22}^{(a2)} Y_{11} R_{12}^{(a1)T} - \tilde{F}_{21},$$

$$M_{22} = R_{22}^{(a2)} Y_{11} R_{22}^{(a1)T} - \tilde{F}_{22}.$$

In cases where $R_{11}^{(a1)}$, $R_{22}^{(a1)}$, $R_{11}^{(a2)}$, and $R_{22}^{(a2)}$ are square and nonsingular or empty, it is possible to solve uniquely for Y. Since Y_{11} is fixed by the constraints, the submatrix M_{22} is also fixed. The submatrix Y_{12} can be selected such that M_{21} is zero. The submatrix Y_{21} can be selected such that M_{12} is zero. When Y_{12}, Y_{21}, and Y_{11} are determined, the submatrix Y_{22} can be selected such that M_{11} is zero. Finally, Y_{11}, Y_{12}, Y_{21}, and Y_{22} are inserted into Y and the problem (2.1) is solved by $X = Q_{b2}^T Y Q_{b1}$. The reason why $\hat{A}_1$ and $\hat{A}_2$ are QR factorized, instead of $\tilde{A}_1$ and $\tilde{A}_2$, is that it had not been equally simple to solve for Y_{12}, Y_{21}, and Y_{22} if the QR factorizations of $\tilde{A}_1$ and $\tilde{A}_2$ had been used.

3. Cases where the solution is not unique. In cases where $R^{(a1)}$ or $R^{(a2)}$ has zero entries in the diagonal the solution of (1.1) is not unique. Also, if $R_{22}^{(a1)}$ or $R_{22}^{(a2)}$ has more columns than rows, then the solution is not unique. Otherwise, the solution is unique. Note that if $m_{b1} = n_{b1}$, then $R_{11}^{(a1)}$, $R_{12}^{(a1)}$, M_{11}, M_{21}, Y_{22}, and

Y_{12} are empty matrices and the solution can be unique even if $m_{a2} < n_{a2}$ and $R^{(a2)}$ has more columns than rows. However, if $m_{b1} < n_{b1}$, then the solution is unique only if all columns of $R^{(a2)}$ are linearly independent. It is the same as if all columns of A_2 are linearly independent. Similarly, if $m_{b2} = n_{b2}$, then the solution can be unique even if $m_{a1} < n_{a1}$. If $m_{b2} < n_{b2}$, then the solution is unique only if the columns of A_1 are linearly independent. This can be expressed in the following way.

THEOREM 3.1. *The constrained least squares problem* (1.1) *has a unique solution if and only if the following two conditions are satisfied:*

- $m_{b1} < n_{b1}$ *and all columns of* A_2 *are linearly independent or* $m_{b1} = n_{b1}$ *and the* $n_{b2} - m_{b2}$ *last columns of* $\tilde{A}_2$ *are linearly independent, where* $\tilde{A}_2 = A_2 Q^T$ *and* Q *is given by the QR factorization of* B_2^T.
- $m_{b2} < n_{b2}$ *and all columns of* A_1 *are linearly independent or* $m_{b2} = n_{b2}$ *and the* $n_{b1} - m_{b1}$ *last columns of* $\tilde{A}_1$ *are linearly independent, where* $\tilde{A}_1 = A_1 Q^T$ *and* Q *is given by the QR factorization of* B_1^T.

Consider the example $m_{a2} = 4$, $m_{a1} = n_{a1} = n_{a2} = n_{b1} = n_{b2} = 5$, $m_{b1} = 3$, $m_{b2} = 2$. In this example $m_{a1}m_{a2} + m_{b1}m_{b2} = 26 > 25 = n_{b1}n_{b2}$. However, the solution is not unique since $R_{22}^{(a2)}$ has more columns than rows ($\in R^{3\times 2}$). More generally, if $m_{a2} < n_{a2}$ and $m_{b1} < n_{b1}$, we always get more than one solution. If $m_{a2} + m_{b2} < n_{a2}$, the submatrix $R_{11}^{(a2)}$ has more columns than rows, and hence the solution is not unique. Similarly, we do not get unique solutions when the inequality $m_{a1} + m_{b1} < n_{b1}$ is satisfied or the inequalities $m_{b2} < n_{b2}$ and $m_{a1} < n_{a1}$ are satisfied. These cases are summarized in the following corollary.

COROLLARY 3.2. *If any of the following four conditions is satisfied, then the solution of* (1.1) *is not unique.*

- $m_{a1} + m_{b1} < n_{a1}$,
- $m_{a2} + m_{b2} < n_{a2}$,
- $m_{b1} < n_{b1}$ *and* $m_{a2} < n_{a2}$,
- $m_{b2} < n_{b2}$ *and* $m_{a1} < n_{a1}$.

4. Analysis using the GSVD. Assume that $m_{a1} \geq n_{a1}$ and $m_{a2} \geq n_{a2}$; then an alternative and more concise analysis of the problem can be derived using the GSVD [1, pp. 157–158], [5, p. 471]. Assume that $\begin{bmatrix} A_1 \\ B_1 \end{bmatrix}$ and $\begin{bmatrix} A_2 \\ B_2 \end{bmatrix}$ are of full column rank. Let the GSVD of (A_1, B_1) be

$$\begin{bmatrix} A_1 \\ B_1 \end{bmatrix} = \begin{bmatrix} U_1 & \\ & V_1 \end{bmatrix} \begin{bmatrix} I & & & \\ & C_1 & & \\ & & & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & & & \\ & S_1 & & \\ & & & I \end{bmatrix} X_1,$$

and let the GSVD of (A_2, B_2) be

$$\begin{bmatrix} A_2 \\ B_2 \end{bmatrix} = \begin{bmatrix} U_2 & \\ & V_2 \end{bmatrix} \begin{bmatrix} I & & & \\ & C_2 & & \\ & & & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & & & \\ & S_2 & & \\ & & & I \end{bmatrix} X_2.$$

Then

$$\|A_2XA_1^T - F\|_F = \left\| U_2 \begin{bmatrix} I & & \\ & C_2 & \\ & & 0 \end{bmatrix} X_2 X X_1^T \begin{bmatrix} I & & \\ & C_1 & \\ & & 0 \end{bmatrix} U_1^T - F \right\|_F$$

$$= \left\| \begin{bmatrix} I & & \\ & C_2 & \\ & & 0 \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} \begin{bmatrix} I & & \\ & C_1 & \\ & & 0 \end{bmatrix} - \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix} \right\|_F,$$

where

$$\begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} = X_1 X X_2^T, \quad \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ F_{21} & F_{22} & F_{23} \\ F_{31} & F_{32} & F_{33} \end{bmatrix} = U_2^T F U_1,$$

and all the matrix partitions are compatible. Similarly

$$B_1 X B_2^T = G$$

gives

$$\begin{bmatrix} 0 & & \\ & S_2 & \\ & & I \end{bmatrix} \begin{bmatrix} X_{11} & X_{12} & X_{13} \\ X_{21} & X_{22} & X_{23} \\ X_{31} & X_{32} & X_{33} \end{bmatrix} \begin{bmatrix} 0 & & \\ & S_1 & \\ & & I \end{bmatrix} = \begin{bmatrix} G_{11} & G_{12} & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & G_{32} & G_{33} \end{bmatrix},$$

where

$$\begin{bmatrix} G_{11} & G_{12} & G_{13} \\ G_{21} & G_{22} & G_{23} \\ G_{31} & G_{32} & G_{33} \end{bmatrix} = U_2^T G U_1.$$

The constraint is consistent if and only if G_{11}, G_{12}, G_{13}, G_{21}, and G_{31} are all zero matrices or empty. The part $\left[\begin{smallmatrix} X_{22} & X_{23} \\ X_{32} & X_{33} \end{smallmatrix}\right]$ is uniquely determined by the constraints and will in the minimization part be a constant matrix. Minimal solution is given by $X_{11} = F_{11}$, $X_{12} = F_{12}C_1^{-1}$, $X_{21} = C_2^{-1}F_{21}$, and X_{13}, X_{31} are arbitrary. It is easily seen that when the column dimension of X_{13} and the row dimension of X_{31} are zero, there is an unique solution. This is satisfied if $\operatorname{rank}\left[\begin{smallmatrix} A_1 \\ B_1 \end{smallmatrix}\right] = \operatorname{rank}(A_1)$ and $\operatorname{rank}\left[\begin{smallmatrix} A_2 \\ B_2 \end{smallmatrix}\right] = \operatorname{rank}(A_2)$. This is one of the cases where the conditions in Theorem 3.1 are satisfied.

The advantage of this GSVD approach is that it is more concise and simple than the approach in sections 2 and 3. However, it takes more computer time to compute two GSVDs than to solve the problem with the ideas in section 2 and check the conditions in Theorem 3.1. Another disadvantage is that the GSVD is not defined when $m_{a1} < n_{a1}$ or $m_{a2} < n_{a2}$. In certain cases where $m_{a1} < n_{a1}$ or $m_{a2} < n_{a2}$, the problem (1.1) can be solved with the ideas in section 2.

5. Complexity analysis. Let us now examine how many flops are required with the method proposed in section 2. We limit the discussion to cases where the solution is unique, $m_{a1} \le n_{a1} = n_{b1} \le m_{b1}$ and $m_{a2} \le n_{a2} = n_{b2} \le m_{b2}$. In other cases, the discussion below gives an overestimate of how many flops are necessary to get

one solution. We use the new standard for flops [5, p. 19], where one flop is one multiplication, one addition, one subtraction, or one division. A QR factorization of an $m \times n$ matrix requires approximately $2n^2(m - n/3)$ flops [5, p. 212]. Hence, to LQ factorize B_1 and B_2 in the first step, $2m_{b1}^2(n_{b1} - m_{b1}/3) + 2m_{b2}^2(n_{b2} - m_{b2}/3)$ flops are required. Then Q_{b1} and Q_{b2} will not be explicitly formed. We do not generate Q_{b1} and Q_{b2}. Instead, we store the Householder multipliers, so that the Householder updates can be carried out as needed, without explicit formation of the orthogonal matrices. To get Y_{11} we have to solve m_{b1} lower triangular systems of dimension $m_{b2} \times m_{b2}$ and m_{b2} lower triangular systems of dimension $m_{b1} \times m_{b1}$. This requires $m_{b1}m_{b2}(m_{b1} + m_{b2})$ flops. To compute $\tilde{A}_1$ and $\tilde{A}_2$ we multiply A_1 and A_2 with Householder updates. This requires about $4(m_{b1}m_{a1}n_{a1} + m_{b2}m_{a2}n_{a2})$ flops [5, p. 197]. The QR factorizations of $\tilde{A}_1$ and $\tilde{A}_2$ require $2n_{a1}^2(m_{a1} - n_{a1}/3)+$ $2n_{a2}^2(m_{a2}-n_{a2}/3)$ flops. To compute $\tilde{F}$ the matrix F is changed through Householder updates. This requires about $4n_{a1}n_{a2}(n_{a1}+n_{a2})$ flops. In a practical implementation, Y_{21}, Y_{12}, and Y_{22} can be computed in the following way:

1. Solve $R_{22}^{(a1)}T_1 = F_{12}^T$.
2. $T_2 := T_1^T - R_{12}^{(a2)}Y_{11}$.
3. Solve $R_{11}^{(a2)}Y_{21} = T_2$.
4. $T_3 := Y_{11}R_{12}^{(a1)T}$.
5. Solve $R_{22}^{(a2)}T_4 = F_{21}$.
6. $T_5 := (T_4 - T_3)^T$.
7. Solve $R_{11}^{(a1)}Y_{12}^T = T_5$.
8. $T_6 := F_{11} - R_{12}^{(a2)}T_3$.
9. Solve $R_{11}^{(a1)}T_7 = T_6^T$.
10. $T_8 := T_7^T - R_{12}^{(a2)}Y_{12}$.
11. Solve $R_{11}^{(a2)}T_9 = T_8$.
12. $T_{10} := R_{12}^{(a1)}Y_{21}^T$.
13. Solve $R_{11}^{(a1)}T_{11} = T_{10}$.
14. $Y_{22} := T_9 - T_{11}^T$.

Note that the systems we solve in steps 1, 3, 5, 7, 9, 11, and 13 are upper triangular. An upper triangular $n \times n$ system, with p right-hand sides, can be solved in pn^2 flops. Hence, step 1 requires $(n_{b2}-m_{b2})m_{b1}^2$ flops. Step 2 requires $2m_{b1}m_{b2}(n_{b1} - m_{b1})$ flops, step 3 requires $(n_{b2}-m_{b2})^2m_{b2}$ flops, step 4 requires $2(n_{b1}-m_{b1})m_{b2}^2$ flops, step 5 requires $(n_{b1}-m_{b1})m_{b2}^2$ flops, step 6 requires $(n_{b1}-m_{b1})m_{b2}$ flops, step 7 requires $m_{b2}(n_{b1}-m_{b1})^2$ flops, step 8 requires $2m_{b2}(n_{b1}-m_{b1})(n_{b2}-m_{b2})$ flops, step 9 requires $(n_{b1}-m_{b1})^2(n_{b2}-m_{b2})$ flops, step 10 requires $2(n_{b2}-m_{b2})(n_{b1}-m_{b1})m_{b1}$ flops, step 11 requires $(n_{b1}-m_{b1})(n_{b2}-m_{b2})^2$ flops, step 12 requires $2(n_{b1}-m_{b1})m_{b1}(n_{b2}-m_{b2})$ flops, step 13 requires $(n_{b1}-m_{b1})^2m_{b2}$ flops, and step 14 requires $(n_{b1}-m_{b1})(n_{b2}-m_{b2})$ flops. Finally, $X = Q_{b2}YQ_{b1}^T$ is computed by multiplication with Householder updates. It can be done in approximately $4m_{b1}m_{b2}(m_{b1}+m_{b2})$ flops. The sum of these expressions is a complicated expression. However, in total it is always less than $35N^3$ flops, where $N = \max(m_{a1}, n_{a1}, m_{a2}, n_{a2})$. This should be compared with how many flops we need if we solve (1.1) as a dense problem, which would require roughly $\mathcal{O}(N^6)$ flops.

6. Summary and conclusions. We have presented a null space method to solve constrained least squares problems of the form (1.1), where $n_{a1} = n_{b1}$, $n_{a2} = n_{b2}$. After transforming the system to the form (2.2), it was trivial to take advantage of the Kronecker product structure in the constraints. The idea to use the QR factorization

of $\hat{A}_1$ and $\hat{A}_2$ (instead of, for instance, the QR factorizations of $\tilde{A}_1$ and $\tilde{A}_2$) makes it possible to also take significant advantage of the Kronecker product structure of the $(A_1 \otimes A_2)$ part. It is also possible to state least squares problems of the form (1.1), where $n_{a1}n_{a2} = n_{b1}n_{b2}$ but $n_{a1} \neq n_{b1}$, $n_{a2} \neq n_{b2}$. In such cases it is not possible to use all the presented ideas, since it is not possible to rewrite it to the form (2.1). Therefore, the "exploded view" (1.1) has some advantage of greater generality over the "compact view" (2.1). However, we only know one application [3] in which we naturally have $n_{a1} = n_{b1}$, $n_{a2} = n_{b2}$. Therefore, we have not made any further analysis of the case $n_{a1} \neq n_{b1}$, $n_{a2} \neq n_{b2}$. We have also shown how the GSVD can be used on these problems and examined the complexity of the algorithm. The ideas in section 2 can be implemented in practically any imperative programming language. A MATLAB implementation can be obtained from the author by sending an email to abrr@cs.umu.se.

Acknowledgments. The author would like to thank Berit Kvernes for her corrections of the English in the paper. I would also like to thank Ji-guang Sun for his suggestions. Finally, I would like to thank the referees for several important comments. Section 4 is due to the comments of one of the referees.

REFERENCES

[1] Å. BJÖRCK, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.
[2] D. W. FAUSETT AND C. T. FULTON, *Large least squares problems involving Kronecker products*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 219–227.
[3] D. R. FORSEY AND R. H. BARTELS, *Surface fitting with hierarchical splines*, ACM Trans. Graphics, 14 (1995), pp. 134–161.
[4] A. GRAHAM, *Kronecker Products and Matrix Calculus with Applications*, Halsted Press, New York, 1981.
[5] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[6] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[7] H. ZHA, *Comments on large least squares problems involving Kronecker products*, SIAM J. Matrix Anal. Appl., 16 (1995), p. 1172.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 161–181, January 1998

011

COMPUTING A FACTOR OF A POLYNOMIAL BY MEANS OF MULTISHIFT *LR* ALGORITHMS*

LUCA GEMIGNANI†

Abstract. In this paper we deal with the numerical approximation of a factor of a polynomial. Our approach is based on the relations between matrix transforms and functional iterations. We show that a generalized LR algorithm applied to an $n \times n$ Hessenberg matrix A may be viewed in a polynomial setting as an iterative method for the computation of a single factor of arbitrary degree $k < n$ of the characteristic polynomial of A. In its basic form our method is linearly convergent under very mild assumptions. The convergence rate can be improved by considering the technique of shifting; the local convergence of our method complemented with a suitable shift strategy is typically quadratic. One iteration of the resulting algorithm can be performed at the overall cost of $O(k^4+nk^3)$ arithmetical operations and $nk^2 \log p/p$ parallel steps with order-pk^2 processors; therefore, it appears to have nice computational features in the typical case where k is a prespecified integer of modest size with respect to n. Moreover, it can be arranged to produce highly efficient parallel algorithms because of its possibility of extensive vectorization. Finally, we confirm its effectiveness by means of numerical experiments which are reported and discussed.

Key words. factorization of polynomials, eigenvalue computation, LR algorithms

AMS subject classifications. 65H05, 65F15

PII. S0895479894277442

1. Introduction.

In this paper we consider the problem of the numerical approximation of a factor of degree k of a polynomial $p(t)$ of degree n. Throughout this discussion, the integer k should be thought of as a fixed prespecified integer of modest size with respect to n. This topic appears, for instance, in the theory of the simultaneous computation of all the zeros of a polynomial in the presence of clusters. In this case, root-finding algorithms usually present a very slow linear convergence and so we expect to improve the convergence behavior by turning on the numerical factorization [5], [16]. Furthermore, the requirement of effective numerical algorithms for the approximate factorization (over the complex field) of a univariate polynomial is also motivated by the observation that in the practice of computation, the coefficients are frequently available only within certain truncation errors, and then the problem of computing the factorization is better conditioned than the zero-finding one.

Our approach relies on the relations between matrix transforms and functional iterations. It is in fact a standard approach to compute the roots of a given polynomial $p(t)$ by computing the eigenvalues of an $n \times n$ Hessenberg matrix A having $p(t)$ as characteristic polynomial. The LR iteration applied to a starting matrix $A = A_1$ defines a sequence of similar matrices by

$$\begin{aligned} A_s - \sigma_s I &= L_s R_s, \\ A_{s+1} - \sigma_s I &= R_s L_s, \quad s \geq 1, \end{aligned} \tag{1}$$

where L_s is unit lower triangular, R_s is upper triangular, and σ_s is a scalar called the shift parameter. Under suitable conditions the sequence $\{A_s\}_{s \in N}$ will tend to upper triangular, or at least block triangular form, yielding information about the

* Received by the editors November 21, 1994; accepted for publication (in revised form) by P. Van Dooren January 14, 1997. This research was supported by funds from the Progetto Analisi numerica e matematica computazionale of MURST and by G.N.I.M. of C.N.R.
http://www.siam.org/journals/simax/19-1/27744.html
† Dipartimento di Informatica, Università di Pisa, 56125 Pisa, Italy (gemi@di.unipi.it).

eigenvalues [20]. Dekker and Traub [6], [7] proved that iteration (1) may be viewed in a functional setting as an iterative scheme to obtain globally convergent algorithms for polynomials. Jenkins and Traub [11], [12] explicitly formed these schemes by obtaining an effective numerical algorithm which has been implemented in the NAPAK library. More recently, the search for methods particularly suited to the parallel architectures has been at the bottom of the increasing interest in the multishift implementation of the matrix LR and QR iterations [19], [1], [8]. Watkins and Elsner [19] studied the convergence properties of the following multishift LR iteration:

$$\begin{aligned} p_s(A_s) &= L_s R_s, \\ A_{s+1} &= L_s^{-1} A_s L_s, \quad s \geq 1, \end{aligned} \tag{2}$$

where $p_s(t)$ is a polynomial of arbitrary degree k_s. If we set $p_s(t)$ as the characteristic polynomial of the $k \times k$ trailing principal submatrix of A_s, then the convergence of the iteration generally results in the separation of a trailing principal submatrix of order close to k which provides for the deflation of a factor of $p(t)$.

Now, let us assume that $A = A_1$ is a lower Hessenberg matrix with unit superdiagonal entries, and denote as $\psi_i^{(s)}(t)$ and $\rho_i^{(s)}(t)$, respectively, the characteristic polynomial of the $i \times i$ leading principal submatrix of A_s and the characteristic polynomial of the $i \times i$ trailing principal submatrix of A_s. Moreover, set $\psi_0^{(s)}(t) = \rho_0^{(s)}(t) = 1$. We may then prove that the matrix iteration (2) is equivalent in a polynomial setting to the following iterative schemes:

$$\begin{cases} p_s(t)\psi_0^{(s+1)}(t) = \sum_{i=0}^{k_s} b_{0,i}^{(s)} \psi_i^{(s)}(t); \\ \cdots\cdots\cdots; \\ \cdots\cdots\cdots; \\ p_s(t)\psi_{n-2}^{(s+1)}(t) = b_{n-2,n-2}^{(s)} \psi_{n-2}^{(s)}(t) + b_{n-2,n-1}^{(s)} \psi_{n-1}^{(s)}(t) + g_{n-2}^{(s)}(t)p(t); \\ p_s(t)\psi_{n-1}^{(s+1)}(t) = b_{n-1,n-1}^{(s)} \psi_{n-1}^{(s)}(t) + g_{n-1}^{(s)}(t)p(t), \end{cases} \tag{3}$$

and

$$\begin{cases} p_s(t)\rho_{n-1}^{(s)}(t) = c_{n-1,n-1}^{(s)} \rho_{n-1}^{(s+1)}(t) + f_{n-1}^{(s)}(t)p(t); \\ p_s(t)\rho_{n-2}^{(s)}(t) = c_{n-2,n-2}^{(s)} \rho_{n-2}^{(s+1)}(t) + c_{n-2,n-1}^{(s)} \rho_{n-1}^{(s+1)}(t) + f_{n-2}^{(s)}(t)p(t); \\ \cdots\cdots\cdots; \\ \cdots\cdots\cdots; \\ p_s(t)\rho_0^{(s)}(t) = \sum_{i=0}^{k_s} c_{0,i}^{(s)} \rho_i^{(s+1)}(t), \end{cases} \tag{4}$$

where $c_{i,j}^{(s)}$ and $b_{i,j}^{(s)}$ are suitable scalars and $f_{n-i}^{(s)}(t)$ and $g_{n-i}^{(s)}(t)$ are suitable polynomials of degree $k_s - i$. The first equation of (4) represents a modification of the well-known Sebastião e Silva method for finding a single zero of $p(t)$. The last equality of (3) reduces to the Jenkins and Traub method in the case where $k_s = 1$. The relation of the Jenkins and Traub globally convergent algorithm to the work of Sebastião e Silva, which was already observed in a different perspective in Householder's book [14], is also made clear.

In the stationary case where $p_1(t) = p_s(t)$ for any s, we are able to show the convergence of the iterative schemes (3) and (4) under very mild assumptions. Specifically, if we assume that

$$|p_1(t_1)| \geq |p_1(t_2)| \geq \cdots \geq |p_1(t_{n-k})| > |p_1(t_{n-k+1})| \geq \cdots |p_1(t_n)| > 0,$$

where t_i denote the eigenvalues of A_1, that is, the zeros of $p(t)$, then for almost any starting matrix A_1 the iterative schemes (3) and (4) can be constructed for any s. Moreover, we have that

$$\left\| \rho_k^{(s+1)}(t) - \prod_{i=n-k+1}^{n} (t - t_i) \right\|_\infty = O(\epsilon^{s+1})$$

and

$$\left\| \psi_{n-k}^{(s+1)}(t) - \prod_{i=1}^{n-k} (t - t_i) \right\|_\infty = O(\epsilon^{s+1}),$$

where ϵ is any number satisfying

$$|p_1(t_{n-k+1})/p_1(t_{n-k})| < \epsilon < 1.$$

In this way, the iterative schemes (3) and (4) provide the means for developing a numerical method for the approximation of a factor of degree $k < n$ of $p(t)$ given a starting approximation $p_1(t)$. The computation of the characteristic polynomials $\psi_{n-i}^{(s+1)}(t)$, $1 \le i \le k$, by using the last k equalities of (3) can be performed in a stable way at the cost of $O(k^4 + nk^3)$ arithmetical operations and $O(nk)$ storage. In a parallel model of computation we easily obtain the upper bound $O(nk^2 \log p/p)$ parallel steps with order-pk^2 processors. According to Brent's scheduling principle for parallel computing, we may decrease the number of processors by a factor s by slowing down parallel computations by $O(s)$ times. On the contrary, no comparable reduction both in cost and in storage can be achieved for the algorithms based on the eigenvalue computation in the case $k \ll n$. Furthermore, as the matrix formulation of (3) and (4) suggests, we can improve the convergence rate by considering a suitable shift strategy. After m steps we replace $p_1(t)$ with the polynomial $\eta_k^{(s)}(t)$ defined as a quotient by the division of $p(t)$ by $\psi_{n-k}^{(s)}(t)$. For an appropriate selection of the parameter m, then we find that the local convergence of the resulting algorithm is typically quadratic.

The iterative schemes (3) and (4) address some of the stability problems of the LR iterations (1) and (2) by avoiding the explicit computation of the entries which define the triangular decomposition. Despite this, the occurrence of breakdowns or near breakdowns in the LR iteration (1) can most severely impact the accuracy of the computed approximations $\psi_{n-i}^{(s+1)}(t)$. Namely, very small changes in the coefficients of the polynomials $\psi_{n-i}^{(s)}(t)$ could in fact lead to substantially larger changes in the coefficients of the polynomials $\psi_{n-i}^{(s+1)}(t)$. However, since the updating procedure employs the original polynomial $p(t)$, poor results produced at a certain step might be corrected in practice in the successive iterations at the cost of increasing the number of iterations.

The paper is organized as follows. In section 2 we give the background on the LR iteration and we describe the basic properties of the numerical schemes which yield its polynomial representation. Section 3 presents a numerical implementation of our algorithm for the numerical approximation of a factor of a polynomial. We confirm its effectiveness by means of numerical experiments which are reported and discussed.

2. Methods. Let $p(t)$ be a monic complex polynomial of degree n such that

$$p(t) = \sum_{i=0}^{n-1} p_i t^i + t^n = \prod_{i=1}^{n} (t - t_i).$$

Let us consider an $n \times n$ lower Hessenberg matrix with unit superdiagonal entries

$$A_1 = \begin{pmatrix} a_{1,1}^{(1)} & 1 & & & \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ a_{n-1,1}^{(1)} & a_{n-1,2}^{(1)} & \cdots & \cdots & 1 \\ a_{n,1}^{(1)} & a_{n,2}^{(1)} & \cdots & \cdots & a_{n,n}^{(1)} \end{pmatrix} \tag{5}$$

such that

$$p(t) = \det(tI - A_1). \tag{6}$$

The standard LR algorithm with explicit shift [15] defines a sequence of similar matrices by

$$\begin{aligned} A_s - \sigma_s I &= L_s R_s, \\ A_{s+1} &= \sigma_s I + R_s L_s, \quad s \geq 1, \end{aligned} \tag{7}$$

where L_s is unit lower triangular, R_s is upper triangular, and σ_s is some shift of the origin of the spectrum of A_s.

Watkins and Elsner [19] proposed the following generalization of (7):

$$\begin{aligned} p_s(A_s) &= L_s R_s, \\ A_{s+1} &= L_s^{-1} A_s L_s, \quad s \geq 1, \end{aligned} \tag{8}$$

where $p_s(t)$ is a monic polynomial of degree k_s less than n. Writing $p_s(t)$ in factored form, we find that the scheme (8) corresponds with k_s steps of (7) where the shifts are the roots of $p_s(t)$. In this way, a representation of the multishift scheme (8) in a polynomial setting can be obtained by considering a generic step of (7).

Let us define the polynomials $\psi_i^{(s)}(t)$ and $\rho_i^{(s)}(t)$ by means of the following equations:

$$\psi_i^{(s)}(t) = \det(tI - \hat{A}_{s,i}), \quad 1 \leq i \leq n-1, \tag{9}$$

and

$$\rho_i^{(s)}(t) = \det(tI - A_{s,i}), \quad 1 \leq i \leq n-1, \tag{10}$$

where

$$A_s = \begin{pmatrix} \hat{A}_{s,n-k} & C_{s,k} \\ B_{s,k} & A_{s,k} \end{pmatrix}$$

with $\hat{A}_{s,n-k} \in \mathbf{C}^{(n-k)\times(n-k)}$ and $A_{s,k} \in \mathbf{C}^{k\times k}$.

The polynomial vector $(\psi_0^{(s)}(t),\dots,\psi_{n-1}^{(s)}(t))^T$, $\psi_0^{(s)}(t)=1$, satisfies the equation

$$t\begin{pmatrix}\psi_0^{(s)}(t)\\ \vdots\\ \psi_{n-1}^{(s)}(t)\end{pmatrix}=A_s\begin{pmatrix}\psi_0^{(s)}(t)\\ \vdots\\ \psi_{n-1}^{(s)}(t)\end{pmatrix}+\begin{pmatrix}0\\ \vdots\\ 0\\ p(t)\end{pmatrix}. \tag{11}$$

Similarly, given a polynomial vector $(\psi_0^{(s)}(t),\dots,\psi_{n-1}^{(s)}(t))^T$ such that $\psi_i^{(s)}(t)$ is a monic polynomial of degree i, there is a unique lower Hessenberg matrix A_s satisfying relation (11) [13]. By replacing the matrix A_s with the matrix JA_s^TJ, where J denotes the $n\times n$ permutation matrix having unit antidiagonal entries, we find that the polynomial vector $(\rho_0^{(s)}(t),\dots,\rho_{n-1}^{(s)}(t))^T$, $\rho_0^{(s)}(t)=1$, is such that

$$t\begin{pmatrix}\rho_{n-1}^{(s)}(t)\\ \vdots\\ \rho_0^{(s)}(t)\end{pmatrix}^T=\begin{pmatrix}\rho_{n-1}^{(s)}(t)\\ \vdots\\ \rho_0^{(s)}(t)\end{pmatrix}^T A_s+\begin{pmatrix}p(t)\\ 0\\ \vdots\\ 0\end{pmatrix}^T. \tag{12}$$

In this way, since we have

$$t\begin{pmatrix}\psi_0^{(s+1)}(t)\\ \vdots\\ \psi_{n-1}^{(s+1)}(t)\end{pmatrix}=L_s^{-1}A_sL_s\begin{pmatrix}\psi_0^{(s+1)}(t)\\ \vdots\\ \psi_{n-1}^{(s+1)}(t)\end{pmatrix}+\begin{pmatrix}0\\ \vdots\\ 0\\ p(t)\end{pmatrix}, \tag{13}$$

we find that

$$\begin{pmatrix}\psi_0^{(s+1)}(t)\\ \vdots\\ \psi_{n-1}^{(s+1)}(t)\end{pmatrix}=L_s^{-1}\begin{pmatrix}\psi_0^{(s)}(t)\\ \vdots\\ \psi_{n-1}^{(s)}(t)\end{pmatrix}.$$

By replacing A_s in (11) with $\sigma_sI+L_sR_s$, we obtain

$$(t-\sigma_s)\begin{pmatrix}\psi_0^{(s+1)}(t)\\ \vdots\\ \psi_{n-1}^{(s+1)}(t)\end{pmatrix}=R_s\begin{pmatrix}\psi_0^{(s)}(t)\\ \vdots\\ \psi_{n-1}^{(s)}(t)\end{pmatrix}+\begin{pmatrix}0\\ \vdots\\ 0\\ p(t)\end{pmatrix}.$$

According to the Hessenberg form of A_s, then R_s is an upper bidiagonal matrix with unit superdiagonal entries

$$R_s=\begin{pmatrix}r_1^{(s)} & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1\\ & & & r_n^{(s)}\end{pmatrix}. \tag{14}$$

This means that the sth step of (7) defines the following scheme:

$$\begin{cases}(t-\sigma_s)\psi_0^{(s+1)}(t)=r_1^{(s)}\psi_0^{(s)}(t)+\psi_1^{(s)}(t);\\ \cdots\cdots\cdots;\\ \cdots\cdots\cdots;\\ (t-\sigma_s)\psi_{n-1}^{(s+1)}(t)=r_n^{(s)}\psi_{n-1}^{(s)}(t)+p(t),\end{cases} \tag{15}$$

which gives

$$\begin{cases} (t-\sigma_s)\psi_0^{(s+1)}(t) = -(\psi_1^{(s)}(\sigma_s)/\psi_0^{(s)}(\sigma_s))\psi_0^{(s)}(t) + \psi_1^{(s)}(t); \\ \cdots\cdots\cdots\cdots; \\ \cdots\cdots\cdots\cdots; \\ (t-\sigma_s)\psi_{n-1}^{(s+1)}(t) = -(p(\sigma_s)/\psi_{n-1}^{(s)}(\sigma_s))\psi_{n-1}^{(s)}(t) + p(t). \end{cases} \tag{16}$$

In the same manner we obtain the following transformation rules for the polynomial vector $(\rho_0^{(s)}(t), \ldots, \rho_{n-1}^{(s)}(t))^T$:

$$\begin{cases} (t-\sigma_s)\rho_{n-1}^{(s)}(t) = r_1^{(s)}\rho_{n-1}^{(s+1)}(t) + p(t); \\ \cdots\cdots\cdots; \\ \cdots\cdots\cdots; \\ (t-\sigma_s)\rho_0^{(s)}(t) = r_n^{(s)}\rho_0^{(s+1)}(t) + \rho_1^{(s+1)}(t). \end{cases} \tag{17}$$

By performing k_s steps of (15) and (17), it can be easily seen that the generalized LR iteration (8) is equivalent in a polynomial setting to the following iterative schemes:

$$\begin{cases} p_s(t)\psi_0^{(s+1)}(t) = \sum_{i=0}^{k_s} b_{0,i}^{(s)}\psi_i^{(s)}(t); \\ \cdots\cdots\cdots; \\ \cdots\cdots\cdots; \\ p_s(t)\psi_{n-2}^{(s+1)}(t) = b_{n-2,n-2}^{(s)}\psi_{n-2}^{(s)}(t) + b_{n-2,n-1}^{(s)}\psi_{n-1}^{(s)}(t) + g_{n-2}^{(s)}(t)p(t); \\ p_s(t)\psi_{n-1}^{(s+1)}(t) = b_{n-1,n-1}^{(s)}\psi_{n-1}^{(s)}(t) + g_{n-1}^{(s)}(t)p(t), \end{cases} \tag{18}$$

and

$$\begin{cases} p_s(t)\rho_{n-1}^{(s)}(t) = c_{n-1,n-1}^{(s)}\rho_{n-1}^{(s+1)}(t) + f_{n-1}^{(s)}(t)p(t); \\ p_s(t)\rho_{n-2}^{(s)}(t) = c_{n-2,n-2}^{(s)}\rho_{n-2}^{(s+1)}(t) + c_{n-2,n-1}^{(s)}\rho_{n-1}^{(s+1)}(t) + f_{n-2}^{(s)}(t)p(t); \\ \cdots\cdots\cdots; \\ \cdots\cdots\cdots; \\ p_s(t)\rho_0^{(s)}(t) = \sum_{i=0}^{k_s} c_{0,i}^{(s)}\rho_i^{(s+1)}(t), \end{cases} \tag{19}$$

where $c_{i,j}^{(s)}$ and $b_{i,j}^{(s)}$ are suitable scalars and $f_{n-i}^{(s)}(t)$ and $g_{n-i}^{(s)}(t)$ are suitable monic polynomials of degree $k_s - i$.

The first equation of (19) defines the well-known Sebastião e Silva method for finding a single zero of $p(t)$. Stewart [18] gave a systematic theoretical description of this method. The last equation of (18) leads to a suitable modification of the Jenkins and Traub globally convergent algorithm for computing a single zero of a polynomial [11], [12]. The relation of Jenkins and Traub algorithm to the work of Sebastião e Silva was already observed in a different perspective in Householder's book [14].

In the basic version of the iterative schemes (18) and (19), we may choose a starting shift $p_1(t)$ of degree $k_1 = k$ and, then, we may set $p_s(t) = p_1(t)\ \forall s \geq 1$; we will refer to this case as the stationary case. If $p_1(t)$ separates h zeros of $p(t)$ from the remaining ones, namely,

$$|p_1(t_1)| \geq |p_1(t_2)| \geq \cdots \geq |p_1(t_h)| > |p_1(t_{h+1})| \geq \cdots |p_1(t_n)| > 0, \tag{20}$$

where

$$\frac{|p_1(t_{h+1})|}{|p_1(t_h)|} \ll \frac{|p_1(t_{h+j+1})|}{|p_1(t_{h+j})|}, \quad 1 \leq j \leq n-h-1,$$

then the convergence of the matrix iterations generally results in the separation of a trailing principal submatrix of order close to $n-h$ which provides for the deflation of a factor of $p(t)$. This suggests to us to consider the following stationary process, called stationary iteration, as a means of approximating a factor of $p(t)$.

Stationary iteration.

Let $p_1(t)$ be a polynomial of degree $k < n$ satisfying (20); moreover, let $\psi_{n-j}^{(0)}(t)$, $1 \le j \le n-h$, be $n-h$ monic polynomials such that $\psi_{n-j}^{(0)}(t)$ has degree $n-j$;

for $s = 1, 2, \ldots$

compute the polynomials $\psi_{n-j}^{(s)}(t)$, $1 \le j \le n-h$, by means of the last $n-h$ equalities of (18);

set $p_{s+1}(t) = p_1(t)$.

end

By a mathematical point of view, we may analogously consider an iterative scheme which uses the first $n-h$ equations of (19). However, as the Jenkins and Traub algorithm has better numerical behavior with respect to the Sebastião e Silva method, we prefer a formulation in terms of the polynomials $\psi_{n-j}^{(s)}(t)$.

As we know, Gaussian elimination fails unless the first $n-1$ leading principal submatrices are nonsingular. This means that the matrix iteration (7) and its polynomial formulation (16) can break down. More specifically, they break down for a certain $s \in \mathbf{N}$ if and only if there exists an index j, $0 \le j \le n-1$, such that $\psi_j^{(s)}(\sigma_s) = 0$. In this case the matrix $A_s - \sigma_s I$ has the leading principal minor of order $j+1$ which is zero and, therefore, its LR factorization does not exist. Now, we observe that in the stationary case where the shift is fixed, the coefficients of $\psi_j^{(s)}(t)$ are rational functions of the coefficients both of $\psi_{j+1}^{(s-1)}(t)$ and of $\psi_{j+2}^{(s-1)}(t)$, where we denote $\psi_n^{(i)}(t) = p(t)$ for any i. By induction, we find that the coefficients of $\psi_j^{(s)}(t)$ are rational functions of the coefficients of $\psi_{j+h}^{(0)}(t)$ for $1 \le h \le n-j$. Thus, we finally obtain that $\psi_j^{(s)}(\sigma_s) = 0$ if and only if the starting polynomial vector belongs to a subset of $\mathbf{C}^{n(n-1)/2}$ with zero Lebesgue measure. Roughly speaking, for almost any starting matrix A_1 or, equivalently, for almost any choice of the polynomials $\psi_j^{(0)}(t)$, $1 \le j \le n-1$, the polynomial vectors $(\psi_0^{(s+1)}(t), \ldots, \psi_{n-1}^{(s+1)}(t))^T$ and $(\rho_0^{(s+1)}(t), \ldots, \rho_{n-1}^{(s+1)}(t))^T$ can be constructed for any $s \in \mathbf{N}$.

In order to investigate the convergence properties of the stationary iteration we will frequently use the following lemma [10].

LEMMA 2.1. *If A and E are $n \times n$ matrices such that A is nonsingular and $||A^{-1}E||_\infty = r < 1$, then $A+E$ is nonsingular*

$$||(A+E)^{-1} - A^{-1}||_\infty \le \frac{||E||_\infty ||A^{-1}||_\infty^2}{1-r}.$$

We first assume that the polynomial $p(t)$ has n distinct zeros such that the inequalities (20) hold. By evaluating the polynomial vector $(\psi_0^{(s+1)}(t), \ldots, \psi_{n-1}^{(s+1)}(t))^T$ at the zeros t_i of $p(t)$, $1 \le i \le n$, we find that

$$B^{(s)} \begin{pmatrix} \psi_{n-1}^{(s)}(t_i) \\ \vdots \\ \psi_0^{(s)}(t_i) \end{pmatrix} = p_1(t_i) \begin{pmatrix} \psi_{n-1}^{(s+1)}(t_i) \\ \vdots \\ \psi_0^{(s+1)}(t_i) \end{pmatrix},$$

where

$$B^{(s)} = \begin{pmatrix} b^{(s)}_{n-1,n-1} & & & & & & \\ b^{(s)}_{n-2,n-1} & b^{(s)}_{n-2,n-2} & & & & & \\ \vdots & \vdots & \ddots & & & & \\ b^{(s)}_{n-k-1,n-1} & \cdots & \cdots & \ddots & & & \\ & \ddots & \cdots & \cdots & \ddots & & \\ O & & \ddots & & & \ddots & \\ & & & b^{(s)}_{0,k} & \cdots & \cdots & b^{(s)}_{0,0} \end{pmatrix}.$$

Inductively, we obtain

$$\begin{pmatrix} \psi^{(s+1)}_{n-1}(t_i) \\ \vdots \\ \psi^{(s+1)}_{0}(t_i) \end{pmatrix} = \left(\frac{1}{p_1(t_i)}\right)^{s+1} \begin{pmatrix} d^{(s)}_{n-1,n-1} & & & \\ d^{(s)}_{n-2,n-1} & d^{(s)}_{n-2,n-2} & & \\ \vdots & \vdots & \ddots & \\ d^{(s)}_{0,n-1} & \cdots & & d^{(s)}_{0\ 0} \end{pmatrix} \begin{pmatrix} \psi^{(0)}_{n-1}(t_i) \\ \vdots \\ \psi^{(0)}_{0}(t_i) \end{pmatrix}. \tag{21}$$

By using the Newton interpolation formula with nodes t_i, $1 \le i \le h$, we can write

$$\psi^{(s+1)}_h(t) - \prod_{i=1}^{h}(t-t_i) = \sum_{i=1}^{h} \psi^{(s+1)}_h[t_1,\dots,t_i] \prod_{j=1}^{i-1}(t-t_j), \tag{22}$$

where $\psi^{(s+1)}_h[t_1,\dots,t_i]$ denotes the $(i-1)$st order divided difference of $\psi^{(s+1)}_h(t)$. We recall that the following formula holds:

$$\psi^{(s+1)}_h[t_1,\dots,t_i] = \sum_{j=1}^{i} \frac{\psi^{(s+1)}_h(t_j)}{\prod^{i}_{l=1,l\neq j}(t_j-t_l)}. \tag{23}$$

Thus, by evaluating (22) at $t = t_m$, $h+1 \le m \le n$, there follows that

$$\psi^{(s+1)}_h(t_m) = \sum_{i=1}^{h} \left\{ \sum_{j=1}^{i} \frac{\psi^{(s+1)}_h(t_j)}{\prod^{i}_{l=1,l\neq j}(t_j-t_l)} \right\} \prod_{j=1}^{i-1}(t_m-t_j) + \prod_{i=1}^{h}(t_m-t_i),$$

which gives

$$\psi^{(s+1)}_h(t_m) = \sum_{i=1}^{h} \psi^{(s+1)}_h(t_i) q_{i,m} + \prod_{i=1}^{h}(t_m-t_i), \tag{24}$$

where the coefficients $q_{i,m}$ are rational expressions depending on the zeros of $p(t)$ only. From (21) there follows that

$$\psi^{(s+1)}_h(t_i) = (1/p_1(t_i))^{s+1} \sum_{l=1}^{n-h} d^{(s)}_{h,n-l} \psi^{(0)}_{n-l}(t_i), \quad 1 \le i \le n. \tag{25}$$

By replacing (25) in (24) we obtain the equalities

$$\left(\frac{1}{p_1(t_m)}\right)^{s+1} \sum_{l=1}^{n-h} d^{(s)}_{h,n-l} \psi^{(0)}_{n-l}(t_m) - \prod_{i=1}^{h}(t_m-t_i)$$
$$= \sum_{i=1}^{h} \left\{ \left(\frac{1}{p_1(t_i)}\right)^{s+1} \sum_{l=1}^{n-h} d^{(s)}_{h,n-l} \psi^{(0)}_{n-l}(t_i) \right\} q_{i,m},$$

which, for m ranging from $h+1$ to n, define an $(n-h)\times(n-h)$ linear system in the $n-h$ variables $d^{(s)}_{h,n-l}$, $1\le l\le n-h$. Namely, we find that

$$\sum_{l=1}^{n-h} d^{(s)}_{h,n-l}\delta^{(s)}_{m-h,l}=\prod_{i=1}^{h}(t_m-t_i),\quad h+1\le m\le n, \tag{26}$$

where the coefficient matrix $\Delta^{(s)}=(\delta^{(s)}_{i,j})$ is given by

$$\delta^{(s)}_{m-h,l}=\frac{\psi^{(0)}_{n-l}(t_m)}{p_1(t_m)^{s+1}}-\sum_{i=1}^{h}\frac{\psi^{(0)}_{n-l}(t_i)q_{i,m}}{p_1(t_i)^{s+1}}. \tag{27}$$

This means that there exists a constant $K>0$, which depends on $p(t)$ and on the starting polynomials $\psi^{(0)}_{n-i}(t)$ only, such that

$$\delta^{(s)}_{i,j}=\frac{1}{p_1(t_{i+h})^{s+1}}\{\psi^{(0)}_{n-j}(t_{i+h})+\epsilon^{(s)}_{i,j}\},\quad 1\le i,j\le n-h, \tag{28}$$

where

$$|\epsilon^{(s)}_{i,j}|\le K\left|\frac{p_1(t_{i+h})}{p_1(t_h)}\right|^{s+1}.$$

Let us denote as P_{n-h} the $(n-h)\times(n-h)$ matrix which has the i,j entry, $1\le i,j\le n-h$, equal to $\psi^{(0)}_{n-j}(t_{i+h})$. Lemma 2.1 yields the following representation of the solution of the linear system (26):

$$d^{(s)}_{h,n-l}=\sum_{j=1}^{n-h}\left(p_1(t_{h+j})^{s+1}\prod_{i=1}^{h}(t_{h+j}-t_i)\right)\left(\gamma_{l,j}+O\left(\left|\frac{p_1(t_{h+1})}{p_1(t_h)^{s+1}}\right|^{s+1}\right)\right), \tag{29}$$

whenever P_{n-h} is nonsingular and $P^{-1}_{n-h}=(\gamma_{i,j})$.

By replacing (29) in (25), we finally obtain that

$$\left\|\psi^{(s+1)}_h(t)-\prod_{i=1}^{h}(t-t_i)\right\|_\infty=O(|p_1(t_{h+1})/p_1(t_h)|^{s+1}),$$

where we set $||p(t)||_\infty=\max_{0\le i\le n}|p_i|$, $p_n=1$. Similarly, we can show that

$$\left\|\rho^{(s+1)}_{n-h}(t)-\prod_{i=h+1}^{n}(t-t_i)\right\|_\infty=O(|p_1(t_{h+1})/p_1(t_h)|^{s+1}).$$

In this way we arrive at the following convergence result for the stationary iteration.

THEOREM 2.2. *Let $p(t)$ be a polynomial of degree n with n distinct zeros numbered in such a way that*

$$|p_1(t_1)|\ge|p_1(t_2)|\ge\cdots\ge|p_1(t_h)|>|p_1(t_{h+1})|\ge\cdots|p_1(t_n)|>0,$$

where $p_1(t)$ is a given polynomial of degree $k<n$. Then, for almost any choice of the polynomials $\psi^{(0)}_{n-j}(t)$, $1\le j\le n-h$, the stationary iteration doesn't break down for any $s\in\mathbf{N}$. Moreover, we have that

$$\left\|\psi^{(s+1)}_h(t)-\prod_{i=1}^{h}(t-t_i)\right\|_\infty=O(|p_1(t_{h+1})/p_1(t_h)|^{s+1}).$$

In matrix form the above theorem states that A_s tends to a block triangular form and, therefore, it implies the linear convergence of the generalized LR iteration (8) in the stationary case whenever A_1 has n distinct eigenvalues. A suitable modification of the Sebastião e Silva method has been introduced in [3] for the simultaneous approximation of all the zeros of a polynomial. The convergence analysis presented in [3, Proposition 3.1] easily yields a convergence result for the generalized LR iteration (8) when applied to a starting matrix A_1 in tridiagonal form.

Virtually any polynomial has distinct zeros; however, in order to cover all of the cases, we describe below a generalization of Theorem 2.2 which deals with multiple zeros.

For the sake of notational simplicity, assume that (20) is satisfied and, moreover, the zeros t_i, $1 \le i \le n$, of $p(t)$ are such that $t_i \neq t_j$ if and only if $i, j \notin \{h-1, h\}$ and $i, j \notin \{h+1, h+2\}$. Now, (22) still holds when we consider the classical extension of the divided differences to the case in which some of the nodes coincide. Under our assumptions we find that

$$\psi_h^{(s+1)}[t_1, \ldots, t_h] = \sum_{j=1}^{h-2} \frac{\psi_h^{(s+1)}(t_j)}{\prod_{l=1, l\neq j}^{h}(t_j - t_l)} + g_1 \psi_h^{(s+1)}(t_h) + g_2 \psi_h^{(s+1)'}(t_h),$$

where g_1 and g_2 are rational expressions depending only on the zeros of $p(t)$ and, moreover, $p'(t)$ denotes the first derivative of $p(t)$. There follows that

$$\psi_h^{(s+1)}(t_m) = \sum_{i=1}^{h-1} \psi_h^{(s+1)}(t_i)\tilde{q}_{i,m} + \psi_h^{(s+1)'}(t_h)\tilde{q}_{h,m} + \prod_{i=1}^{h}(t_m - t_i), \tag{30}$$

where the coefficients $\tilde{q}_{i,m}$ are rational expressions depending on the zeros of $p(t)$ only and $m = h+2, \ldots, n$. Further, by evaluating $\psi_h^{(s+1)'}(t)$, we find that

$$\psi_h^{(s+1)'}(t_{h+1}) = \sum_{i=1}^{h-1} \psi_h^{(s+1)}(t_i)\tilde{q}_{i,h+1} + \psi_h^{(s+1)'}(t_h)\tilde{q}_{h,h+1} + \tilde{q}_{h+1,h+1}, \tag{31}$$

where the coefficients $\tilde{q}_{i,h+1}$ are rational expressions depending on the zeros of $p(t)$ again. It can be easily seen that (18) and (21) inductively yield the following formula:

$$\psi_h^{(s+1)'}(\bar{t}\,) = \sum_{l=1}^{n-h} d_{h,n-l}^{(s)} \left\{ \frac{\psi_{n-l}^{(0)\,\prime}(\bar{t}\,)}{p_1(\bar{t}\,)^{s+1}} - \frac{\psi_{n-l}^{(0)}(\bar{t}\,)(s+1)p_1{}'(\bar{t}\,)}{p_1(\bar{t}\,)^{s+2}} \right\}, \tag{32}$$

where $\bar{t} = t_h, t_{h+1}$. By replacing (25) and (32) in both (31) and (30), we arrive at an $(n-h)\times(n-h)$ linear system in the $n-h$ variables $d_{h,n-l}^{(s)}$. For the coefficient matrix $\Delta^{(s)} = (\delta_{i,j}^{(s)})$, we now have

$$\begin{aligned}\delta_{1,l}^{(s)} = {} & \frac{\psi_{n-l}^{(0)\,\prime}(t_{h+1})}{p_1(t_{h+1})^{s+1}} - \frac{\psi_{n-l}^{(0)}(t_{h+1})(s+1)p_1{}'(t_{h+1})}{p_1(t_{h+1})^{s+2}} - \sum_{i=1}^{h-1} \frac{\psi_{n-l}^{(0)}(t_i)\tilde{q}_{i,h+1}}{p_1(t_i)^{s+1}} \\ & -\tilde{q}_{h,h+1} \left\{ \frac{\psi_{n-l}^{(0)\,\prime}(t_h)}{p_1(t_h)^{s+1}} - \frac{\psi_{n-l}^{(0)}(t_h)(s+1)p_1{}'(t_h)}{p_1(t_h)^{s+2}} \right\}, \quad 1 \le l \le n-h,\end{aligned}$$

and, moreover,

$$\delta_{m-h,l}^{(s)} = \frac{\psi_{n-l}^{(0)}(t_m)}{p_1(t_m)^{s+1}} - \sum_{i=1}^{h-1} \frac{\psi_{n-l}^{(0)}(t_i)\tilde{q}_{i,m}}{p_1(t_i)^{s+1}} - \tilde{q}_{h,m}\left\{ \frac{\psi_{n-l}^{(0)\prime}(t_h)}{p_1(t_h)^{s+1}} - \frac{\psi_{n-l}^{(0)}(t_h)(s+1)p_1'(t_h)}{p_1(t_h)^{s+2}} \right\},$$

for $h+2 \le m \le n$ and $1 \le l \le n-h$. These equations imply the following estimates:

$$\delta_{1,l}^{(s)} = \frac{1}{p_1(t_{h+1})^{s+1}} \left\{ \psi_{n-l}^{(0)\prime}(t_{h+1}) - \frac{\psi_{n-l}^{(0)}(t_{h+1})(s+1)p_1'(t_{h+1})}{p_1(t_{h+1})} + O\left((s+1) \left| \frac{p_1(t_{h+1})}{p_1(t_h)} \right|^{s+1} \right) \right\},$$

and

$$\delta_{i,l}^{(s)} = \frac{1}{p_1(t_{h+i})^{s+1}} \left\{ \psi_{n-l}^{(0)\prime}(t_{h+i}) + O\left((s+1) \left| \frac{p_1(t_{h+i})}{p_1(t_h)} \right|^{s+1} \right) \right\}.$$

Let $\tilde{P}_{n-h}$ be the $(n-h) \times (n-h)$ matrix which has the i,j entry equal to $\psi_{n-j}^{(0)\prime}(t_{h+1})$ if $i = 1$ and $\psi_{n-j}^{(0)}(t_{h+i})$ otherwise. Moreover, let us denote as $\mathbf{u}_{n-h}$ the vector with entries $\psi_{n-l}^{(0)}(t_{h+1})(s+1)(p_1'(t_{h+1})/p_1(t_{h+1}))$, $1 \le l \le n-h$. Observe that $\mathbf{u}_{n-h}$ is parallel with the second column of $\tilde{P}_{n-h}^T$. Hence, if we assume that $\tilde{P}_{n-h}$ is nonsingular, an application of the Sherman–Morrison formula says that $\tilde{P}_{n-h}^T - \mathbf{u}_{n-h}\mathbf{e}_1^T$ is nonsingular and

$$(\tilde{P}_{n-h}^T - \mathbf{u}_{n-h}\mathbf{e}_1^T)^{-1} = \left(I + \frac{(s+1)p_1'(t_{h+1})}{p_1(t_{h+1})} \mathbf{e}_2\mathbf{e}_1^T \right) \tilde{P}_{n-h}^{-T}.$$

By applying Lemma 2.1, we conclude that

$$\left\| \psi_h^{(s+1)}(t) - \prod_{i=1}^{h}(t-t_i) \right\|_\infty = O((s+1)^2 |p_1(t_{h+1})/p_1(t_h)|^{s+1}).$$

The next theorem generalizes Theorem 2.2 to the case where no assumptions about the multiplicity of the zeros of $p(t)$ are made.

THEOREM 2.3. *Let $p(t)$ be a polynomial of degree n. Assume that its zeros t_i, $1 \le i \le n$, are numbered so that*

$$|p_1(t_1)| \ge |p_1(t_2)| \ge \cdots \ge |p_1(t_h)| > |p_1(t_{h+1})| \ge \cdots |p_1(t_n)| > 0,$$

where $p_1(t)$ is a given polynomial of degree $k < n$. Then, for almost any choice of the polynomials $\psi_{n-j}^{(0)}(t)$, $1 \le j \le n-h$, the stationary iteration doesn't break down for any $s \in \mathbf{N}$. Moreover, we have that

$$\left\| \psi_h^{(s+1)}(t) - \prod_{i=1}^{h}(t-t_i) \right\|_\infty = O(\epsilon^{s+1}),$$

where ϵ is any number satisfying

$$|p_1(t_{h+1})/p_1(t_h)| < \epsilon < 1.$$

Theorems 2.2 and 2.3 say that the stationary iteration can be used in order to construct iterative methods for approximating a factor of $p(t)$ of degree $n-h=k$ whenever a starting approximation $p_1(t)$ of degree k satisfying (20) is available. To be specific, denote as $\{\psi_{n-k}^{(s)}(t)\}_{s\in\mathbf{N}}$ a polynomial sequence generated by the stationary iteration. Introduce the polynomials $\eta_k^{(s)}(t)$ which are defined as a quotient by the division of $p(t)$ by $\psi_{n-k}^{(s)}(t)$, that is,

$$p(t)=\eta_k^{(s)}(t)\psi_{n-k}^{(s)}(t)+\theta^{(s)}(t), \tag{33}$$

where the degree of $\theta^{(s)}(t)$ is less than the degree of $\psi_{n-k}^{(s)}(t)$. In the case where $\psi_{n-k}^{(s)}(t)$ approaches the polynomial $\prod_{i=1}^{n-k}(t-t_i)$, the quotient $\eta_k^{(s)}(t)$ should converge to the polynomial $\prod_{i=n-k+1}^{n}(t-t_i)$. The following iterative process exploits this observation.

Stationary factor iteration.

Let $p_1(t)$ be a polynomial of degree $k<n$ satisfying (20) with $h=n-k$; moreover, let $\psi_{n-j}^{(0)}(t)$, $1\le j\le k$, be k monic polynomials such that $\psi_{n-j}^{(0)}(t)$ has degree $n-j$;

for $s=1,2,\ldots$

compute the polynomials $\psi_{n-j}^{(s)}(t)$, $1\le j\le k$, by means of the last k equalities of (18);

compute $\eta_k^{(s)}(t)$ and check for its convergence;

set $p_{s+1}(t)=p_1(t)$.

end

The convergence of the stationary factor iteration is typically linear as the next theorem shows.

THEOREM 2.4. *Let $p(t)$ be a polynomial of degree n. Assume that its zeros t_i, $1\le i\le n$, are numbered so that*

$$|p_1(t_1)|\ge|p_1(t_2)|\ge\cdots\ge|p_1(t_{n-k})|>|p_1(t_{n-k+1})|\ge\cdots|p_1(t_n)|>0,$$

where $p_1(t)$ is a given polynomial of degree $k<n$. Then, for almost any choice of the polynomials $\psi_{n-j}^{(0)}(t)$, $1\le j\le k$, the stationary factor iteration doesn't break down for any $s\in\mathbf{N}$. Moreover, we have that

$$\left\|\eta_k^{(s+1)}(t)-\prod_{i=1}^{n-k}(t-t_i)\right\|_\infty=O(\epsilon^{s+1}),$$

where ϵ is any number satisfying

$$|p_1(t_{n-k+1})/p_1(t_{n-k})|<\epsilon<1.$$

Proof. By viewing (33) in matrix form, we find that the coefficients of $\eta_k^{(s)}(t)$ solve a linear system. The coefficient matrix is an upper triangular Toeplitz matrix determined by the coefficients of the polynomial $\psi_{n-k}^{(s)}(t)$; the known vector consists of the coefficients of $p(t)$. Hence, the thesis follows by combining the results of Theorem 2.3 and of Lemma 2.1. □

In order to improve the linear convergence rate, the matrix formulation of (18) and (19) suggests to us to consider the shift strategies which have been used with

success in matrix iteration theory. Roughly speaking, Theorem 2.4 says that $\eta_k^{(s)}(t)$ is a good separator of the spectrum of $p(t)$, that is,

$$\frac{|\eta_k^{(s)}(t_{n-k+1})|}{|\eta_k^{(s)}(t_{n-k})|} \ll 1,$$

for all sufficiently large s. This motivates the following modification of the stationary factor iteration which includes a generalized shift strategy.

Shifted factor iteration.

Let $p_1(t)$ be a polynomial of degree $k < n$ satisfying (20) with $h = n-k$; moreover, let $\psi_{n-j}^{(0)}(t)$, $1 \le j \le k$, be k monic polynomials such that $\psi_{n-j}^{(0)}(t)$ has degree $n-j$;

for $s = 1, 2, \ldots$

compute the polynomials $\psi_{n-j}^{(s)}(t)$, $1 \le j \le k$, by means of the last k equalities of (18);

set $p_{s+1}(t) = p_1(t)$ if $0 \le s \le m$. Otherwise, if $s > m$, compute $\eta_k^{(s)}(t)$ as defined by (33) and check for its convergence; then, set $p_{s+1}(t) = \eta_k^{(s)}(t)$.

end

At this point, we do not discuss how to choose the parameter m and we assume that we have a well-defined criterion which says whether $s > m$ or not. We will return to this topic in the next section.

The local convergence of the shifted factor iteration is typically quadratic as stated in the next theorem.

THEOREM 2.5. *Let $p(t)$ be a polynomial of degree n with n distinct zeros numbered so that*

$$|p_1(t_1)| \ge |p_1(t_2)| \ge \cdots \ge |p_1(t_{n-k})| > |p_1(t_{n-k+1})| \ge \cdots |p_1(t_n)| > 0,$$

where $p_1(t)$ is a given polynomial of degree $k < n$. For a given choice of the starting polynomials $\psi_{n-j}^{(0)}(t)$, $1 \le j \le k$, assume that the shifted factor iteration does not break down for any selection of m. Moreover, assume that the $k \times k$ matrix $P_k = (\psi_{n-j}^{(0)}(t_{i+n-k}))$ is nonsingular. Then the local convergence of the shifted factor iteration is quadratic in the following sense: $\exists \bar{n} \in \mathbf{N}$ such that $\forall m \ge \bar{n}$ $\exists$ M_m such that

$$\left\| \eta_k^{(s+1)}(t) - \prod_{i=n-k+1}^{n} (t - t_i) \right\|_\infty \le M_m \left\| \eta_k^{(s)}(t) - \prod_{i=n-k+1}^{n} (t - t_i) \right\|_\infty^2$$

is fulfilled $\forall s > m$.

Proof. Assume that the cumulative polynomial shift $\prod_{l=1}^{s} p_l(t)$ satisfies

$$\frac{\max_{n-k+1 \le j \le n} \prod_{l=1}^{s} |p_l(t_j)|}{\min_{1 \le j \le n-k} \prod_{l=1}^{s} |p_l(t_j)|} \le \epsilon$$

for a sufficiently small $\epsilon > 0$. By viewing the proof of Theorem 2.2, we find that the entries $\delta_{i,j}^{(s-1)}$ of (28) are now defined by

$$\delta_{i,j}^{(s-1)} = \frac{1}{\prod_{l=1}^{s} p_l(t_{i+n-k})} \{\psi_{n-j}^{(0)}(t_{i+n-k}) + \epsilon_{i,j}^{(s-1)}\}, \quad 1 \le i, j \le k,$$

where the small perturbations $\epsilon_{i,j}^{(s-1)}$ are bounded by

$$|\epsilon_{i,j}^{(s)}| \leq K\epsilon$$

for the same constant K. Hence, we have that

$$\left\| \psi_{n-k}^{(s)}(t) - \prod_{i=1}^{n-k}(t-t_i) \right\|_\infty \leq C_1\epsilon,$$

where C_1 is a suitable constant which depends on $p(t)$ and the starting polynomials $\psi_i^{(0)}(t)$. Following the proof of Theorem 2.4, we are able to show that there exists a constant $C_2 > 0$ depending only on $p(t)$ such that

$$\left\| \eta_k^{(s)}(t) - \prod_{i=n-k+1}^{n}(t-t_i) \right\|_\infty \leq C_1C_2\epsilon.$$

In this way, we are able to show that there exists a constant $C_3 > 0$, which depends only on $p(t)$ again, such that

$$|\eta_k^{(s)}(t_{\omega(n-k+1)})| \leq C_1C_2C_3\epsilon,$$

and

$$\left| \eta_k^{(s)}(t_{\sigma(n-k)}) - \prod_{i=n-k+1}^{n}(t_{\sigma(n-k)}-t_i) \right|_\infty \leq C_1C_2C_3\epsilon.$$

Here σ and ω are suitable permutations of $\{1,\ldots,n-k\}$ and $\{n-k+1,\ldots,n\}$, respectively; they are defined by

$$|\eta_k^{(s)}(t_{\sigma(1)}| \geq \cdots \geq |\eta_k^{(s)}(t_{\sigma(n-k)})| > |\eta_k^{(s)}(t_{\omega(n-k+1)})| \geq \cdots |\eta_k^{(s)}(t_{\omega(n)})| \geq 0.$$

Thus, we conclude that

$$\frac{\max_{n-k+1\leq j\leq n}\prod_{l=1}^{s+1}|p_l(t_j)|}{\min_{1\leq j\leq n-k}\prod_{l=1}^{s+1}|p_l(t_j)|} \leq C\epsilon^2,$$

where C is a suitable constant which depends on $p(t)$ and the starting polynomials $\psi_i^{(0)}(t)$. This implies the local quadratic convergence of the shifted factor iteration. □

The shifted factor iteration reduces to the Jenkins–Traub methods [11], [12] in the case where we look for a linear factor ($k = 1$). Some computational experience with the algorithms developed here for the numerical approximation of a factor of a polynomial is the subject of the following section.

3. Computational results. In this section, we discuss a preliminary numerical implementation of the shifted factor iteration for approximating a factor

$$p^\star(t) = \prod_{i=1}^{k}(t-t_{n-k+i})$$

of a monic complex polynomial

$$p(t) = \sum_{i=0}^{n-1} p_i t^i + t^n = \prod_{i=1}^{n} (t - t_i)$$

given a starting approximation $p_1(t)$ of degree k which satisfies the condition (20), namely,

$$|p_1(t_1)| \geq |p_1(t_2)| \geq \cdots \geq |p_1(t_{n-k})| > |p_1(t_{n-k+1})| \geq \cdots |p_1(t_n)| > 0. \tag{34}$$

By a computational point of view, the proposed algorithm has several advantages over more traditional approaches based on the eigenvalue computation. In the case where k is an integer of modest size with respect to n, one step of the resulting algorithm can be performed at a sequential cost which is almost linear with respect to n. The storage needed is also linear with respect to n. Furthermore, our implementation is particularly suited to the parallel architectures because of its possibility of extensive vectorization. One iteration requires $nk^2 \log p/p$ parallel steps with order-pk^2 processors.

On the other hand, there still exist some numerical difficulties which stand in the way of a robust implementation of the methods previously developed. First, the separation between the spectrum of $p^\star(t)$ and the remaining zeros of $p(t)$ seems to affect the conditioning of the factorization problem. In fact, the equation $p(t) = p^\star(t)\hat{p}(t)$ can be seen as a nonlinear equation in the coefficients of the two factors. A linearization technique such as the Newton method requires us to solve a linear system with the Jacobian matrix as the coefficient matrix. In this case the Jacobian matrix at the solution is the resultant of $p^\star(t)$ and $\hat{p}(t)$ which is singular if and only if $p^\star(t)$ and $\hat{p}(t)$ have at least one common zero [9]. Second, the recovery of the coefficients of the polynomials $\psi_i^{(s+1)}(t)$ generated by the *LR* iteration applied to a Hessenberg matrix, which has the polynomials $\psi_i^{(s)}(t)$ as characteristic polynomials of its leading principal submatrices, can be an ill-conditioned problem. More specifically, a very small change in the coefficients of the polynomials $\psi_i^{(s)}(t)$ can in fact lead to substantially larger changes in the coefficients of $\psi_i^{(s+1)}(t)$ whenever a near breakdown occurs in the triangular factorization. However, since the updating procedure employs the original polynomial $p(t)$, poor results produced at a certain step might be corrected in practice in the successive iterations at the cost of increasing the number of iterations.

One step of the shifted factor iteration can be organized into three phases:

1. compute the polynomials $\psi_{n-j}^{(s)}(t)$, $1 \leq j \leq k$;
2. compute the polynomial $\eta_k^{(s)}(t)$ defined by (33);
3. check for the convergence of $\eta_k^{(s)}(t)$ and decide whether to start with the shift strategy.

Concerning the first phase, we express the last k equations of (18) in matrix form. In this way, we are able to reduce the computation of the coefficients of the polynomial $\psi_{n-i}^{(s+1)}(t)$, $1 \leq i \leq k$, to find the solution of a suitable linear system $Ax = b$, where $A \in \mathbf{C}^{(n+k-i)\times(n+k-i)}$. The coefficient matrix A can be partitioned into a block form as follows:

$$A = \begin{pmatrix} T_1 & T_2 \\ T_3 & T_4 \end{pmatrix},$$

where $T_1 \in \mathbf{C}^{k\times k}$ and $T_4 \in \mathbf{C}^{n-i\times n-i}$ is a band upper triangular Toeplitz matrix with bandwidth k. Let

$$b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

be the corresponding partition of the known vector b. In view of the block triangular decomposition

$$A = \begin{pmatrix} I_k & T_2T_4^{-1} \\ 0 & I_{n-i} \end{pmatrix} \begin{pmatrix} T_1 - T_2T_4^{-1}T_3 & 0 \\ T_3 & T_4 \end{pmatrix},$$

we solve the linear system $Ax = b$ by performing the following two steps:

1. we compute the vector

$$\begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix} = \begin{pmatrix} I_k & T_2T_4^{-1} \\ 0 & I_{n-i} \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix};$$

2. we solve the lower block triangular system

$$\begin{pmatrix} T_1 - T_2T_4^{-1}T_3 & 0 \\ T_3 & T_4 \end{pmatrix} x = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix}.$$

Since we have

$$\begin{pmatrix} I_k & T_2T_4^{-1} \\ 0 & I_{n-i} \end{pmatrix}^{-1} = \begin{pmatrix} I_k & -T_2T_4^{-1} \\ 0 & I_{n-i} \end{pmatrix},$$

the first step amounts to solving a linear system with coefficient matrix T_4 and, therefore, it requires $O(nk)$ arithmetical operations or $nk^2 \log p/p$ parallel steps with order-p processors [4]. Concerning the second step, we have firstly to compute the matrix $T_4^{-1}T_3$ by solving k linear systems with coefficient matrix T_4 at the overall cost of $O(nk^2)$ arithmetical operations or $nk^2 \log p/p$ parallel steps with order-pk processors. Then, we finally solve the linear system with coefficient matrix $T_1 - T_2T_4^{-1}T_3$ by determining its QR factorization at the cost of $O(k^3)$ arithmetical operations and k^2 parallel steps with order-k processors. Pivoting and balancing are performed in order to improve the stability of the QR decomposition. Therefore, we compute the coefficients of $\psi_{n-i}^{(s+1)}(t)$ in a stable way at the cost of $O(k^3 + nk^2)$ arithmetical operations or $nk^2 \log p/p$ parallel steps with order-pk processors. This yields the overall cost of $O(k^4+nk^3)$ arithmetical operations or $nk^2 \log p/p$ parallel steps with order-pk^2 processors for executing the first phase of the shifted factor iteration.

The computation of the coefficients of $\eta_k^{(s)}(t)$ defined by (33) amounts to solving a $k \times k$ Toeplitz linear system. This task can be accomplished in a stable way by using $O(k^3)$ arithmetical operations or k^2 parallel steps with order-k processors. In brief, we perform one step of the shifted factor iteration at the overall cost of $O(k^4 + nk^3)$ arithmetical operations or $nk^2 \log p/p$ parallel steps with order-pk^2 processors.

The major decision of our algorithm concerns the selection of the parameter m. The well-known results on the convergence of both the LR algorithm [20] and the Jenkins–Traub algorithms [11], [12] have shown that it is efficient to start with a shift strategy only after a very weak test for the convergence has been passed. However, some difficulties arise when the degree k of the desired factor is greater than two. In fact, for the convergence we should guarantee that the polynomials $\eta_k^{(s+1)}(t)$,

TABLE 1

i	*cond*
1	10^{13}
2	10^{21}
3	10^{17}
4	10^{13}
5	10^{15}

$s > m$, satisfy the condition (34). This means that the selection of the parameter m needs some preliminary information about the root distribution in the complex plane, that is, about the separation between the wanted and unwanted spectrum. Our implementation performs the shift strategy described in the above section when

$$|||\eta_k^{(s+1)}(t)||_\infty - ||\eta_k^{(s)}(t)||_\infty| \leq \ \eta,$$

where, at present, ad hoc choices are made for η. The algorithm is halted when either we find that

$$||\eta_k^{(s+1)}(t) - \eta_k^{(s)}(t)||_\infty \leq \epsilon\ (n+k)||p(t)||_\infty, \tag{35}$$

where ϵ denotes here the machine precision, or the number of iterations exceeds a fixed value $itmax$. In the latter case the program reports failure.

We have implemented our algorithm by using Mathematica$^{\text{TM}}$ on Macintosh with low-precision arithmetic, i.e., about 18 decimal digits. We performed numerical experiments for checking the convergence behavior of our algorithm when either approximations of the coefficients or approximations of the zeros are known. The latter situation is particularly interesting for complementing the root-finding algorithms based on the simultaneous approximation of the roots in the presence of clusters.

The first example consists of the following polynomial which is used as test polynomial for the root-finding methods:

$$p(t) = (t-2)^{10}(t-1.5)\prod_{i=1}^{4}(t-1-a^i),$$

where $a = 0.01$. The classical Wilkinson's analysis about the sensitivity of the polynomial roots indicates that the conditioning of computing a root α, that is, a linear factor, depends on the value $p'(\alpha)$ and, therefore, it depends on the separation between α and the remaining spectrum of $p(t)$. In this case most of the roots of $p(t)$ are ill conditioned. On the other hand, if we measure the conditioning of splitting $p(t)$ in two factors $p^\star(t)$ and $\hat{p}(t)$ by means of the spectral condition number, that is, the ratio of the largest singular value to the smallest one, of the resultant matrix generated by $p^\star(t)$ and $\hat{p}(t)$, then Table 1 shows that no advantage is achieved by passing from the root-finding problem to the factorization one. We denote $p_i^\star(t) = \prod_{j=1}^{i}(t - t_{n-j+1})$, where the zeros t_i of $p(t)$ are numbered so that (34) holds with $p_1(t) = (t-1)^i$. For $i = 1, \ldots, 5$ we compute the spectral condition number *cond* of the resultant matrix generated by $p_i^\star(t)$ and $p(t)/p_i^\star(t)$ by means of the Mathematica$^{\text{TM}}$ function $SingularValues$.

Special attention should be given to the case $i = 5$. Although the spectrum of $p_5^\star(t)$ is well separated from the remaining zeros of $p(t)$, the corresponding factorization

problem is extremely ill conditioned. This resembles what happens for the polynomial roots. We made use of our algorithm in order to approximate the factor $p_5^\star(t)$ of $p(t)$. We considered different choices of η and $itmax$. In all of the cases our program reports failure.

The second example is a suitable modification of the first one. We consider the polynomial

$$p(t) = (t^{10} - 2^{10})(t - 1.5)\prod_{i=1}^{4}(t - 1 - a^i),$$

where $a = 0.01$ and $p_1(t) = (t-1)^5$. The conditioning of factoring $p(t)$ as $p^\star(t)\hat{p}(t)$, where

$$p^\star(t) = (t - 1.5)\prod_{i=1}^{4}(t - 1 - a^i),$$

is now of order 10^6. If we set $\eta = 0.01$ and $itmax = 100$, our implementation reports failure. However, if we denote as $\eta_5^{(s)}(t)$ the computed approximations of $p^\star(t)$, the condition

$$||\eta_5^{(s+1)}(t) - \eta_5^{(s)}(t)||_\infty \leq \ 10^{-11}$$

is satisfied already for $s = 8$. Moreover, we find that

$$||\eta_5^{(9)}(t) - p^\star(t)||_\infty \leq \ 10^{-11}.$$

This means that the stopping criterion (35) is generally rather inadequate. A more reliable one should take into account some information about the conditioning of the problem. How to find this information in a fast way is an open question and its solution seems to need the extension of classical theorems on the sensitivity of polynomial roots to the case where factors of arbitrary degree are considered. Some results on this topic can be found in [17].

The previous examples deal with polynomials which are pathological in a certain sense. The problem of factoring polynomials of low or moderate degree having random complex coefficients uniformly distributed is usually a well-conditioned problem. In order to substantiate our belief numerically, we have used the proposed algorithm for determining an approximate factor of degree 4 of 20 monic complex polynomials of degree 15 with random coefficients $a + ib$, where $i^2 = -1$ and a and b are drawn from the uniform distribution in the interval $[0, 1]$. We set $p_1(t) = (t-1)^4$ and, therefore, we look for the 4 zeros of $p(t)$ which are the nearest to 1. For any polynomial $p(t)$ we compute these zeros by means of the Mathematica$^{\mathrm{TM}}$ function $NRoots$. We multiply the corresponding linear factor and then we refer to the resulting polynomial of degree 4 as to $p^\star(t)$. The spectral condition number $cond$ of the resultant matrix generated by $p^\star(t)$ and $p(t)/p^\star(t)$ is approximately of order 10^3 on average. We initialize the program with $\eta = 0.01$ and $itmax = 100$; moreover, the monic polynomials $\psi_{n-l}^{(0)}(t)$, $1 \leq l \leq 4$, coincide with the normalized derivatives of $p(t)$ of appropriate order. The algorithm reached the modified convergence condition

$$||\eta_4^{(s+1)}(t) - \eta_4^{(s)}(t)||_\infty \leq \ cond\ \epsilon\ (n+4)||p(t)||_\infty$$

TABLE 2

k	$n = 50$	$n = 100$	$n = 150$	$n = 200$
5	6	6	6	7
10	18	19	20	20

in all of the cases. The medium number of iterations is about 16. If we denote as $\eta_4^{(\bar{s})}(t)$ the approximation of $p^\star(t)$ produced as output by the program, then the condition

$$||\eta_4^{(\bar{s})}(t) - p^\star(t)||_\infty \leq \ cond\ \epsilon\ (n+4)^2 ||p(t)||_\infty$$

is always satisfied. As we increase the polynomial degree to 20, the spectral condition number increases to the order of 10^4 on average. Our program reports failure in two cases where the separation ratio $|p_1(t_{17})/p_1(t_{16})|$ is greater than 0.95. Similar results have been obtained for polynomials with random coefficients with normal distribution with mean 0 and standard deviation 1.

Now, to exhibiting behavior on polynomials of high degree, we considered the following two sets of test polynomials.

1. The first set consists of polynomials $p(t)$ such that

$$p(t) = \hat{p}(t)p^\star(t), \quad p^\star(t) = (t-3)^k + 10^{-4}\sum_{j=0}^{k-1} t^j,$$

where $\hat{p}(t)$ is a complex polynomial having randomly generated coefficients with real and imaginary parts between -1 and 1. We considered the cases $k = 5$ and $k = 10$. For each fixed value of k, we generate 100 polynomials $s(t)$ with degrees 50, 100, 150, and 200. The monic polynomials $\psi_{n-j}^{(0)}(t)$, $1 \leq j \leq k$, coincide with the normalized derivatives of $p(t)$ of appropriate order. For each fixed value of k, we use the polynomial

$$p_1(t) = \prod_{j=1}^{k}(t - 3 - 10^{-4}w^j), \quad w = \cos(\pi/k) + i\sin(\pi/k),\ i^2 = -1,$$

as the starting approximation of $p^\star(t)$. We chose $\eta = 10^{-2}$ and $itmax = 50$. In all cases the computation has been carried out successfully and our implementation always reaches the stopping condition (35). This fact is a manifestation of the well conditioning of the considered factorization problem. If we denote as $\eta_k^{(\bar{s})}(t)$ the approximation of $p^\star(t)$ produced as output by the program, then we find that the inequality

$$||\eta_k^{(\bar{s})}(t) - p^\star(t)||_\infty \leq C\epsilon\ (n+k)||p(t)||_\infty$$

is satisfied in all of the cases. Here C is a constant of moderate size, say $C \leq 10$. Table 2 reports the average number of iterations rounded to the nearest integer. This table clearly shows that the convergence properties of our method are completely independent of the degree n of $p(t)$.

2. The second set consists of polynomials $p(t)$ such that

$$p(t) = \hat{p}(t)p^\star(t), \quad p^\star(t) = \prod_{j=-2,\ j\neq 0}^{2}(t - a + 0.1\ j)(t - a + 0.1\ i\ j),$$

TABLE 3

a	$n = 50$	$n = 100$	$n = 150$	$n = 200$
2	6	6	6	6
1.5	16	18	11	11

where $i^2 = -1$ and $\hat{p}(t)$ is a complex polynomial having randomly generated coefficients with real and imaginary parts between -1 and 1. We generate 100 polynomials $\hat{p}(t)$ for each of the following values of the degree n, $n \in \{50, 100, 150, 200\}$. In order to approximate the coefficients of the factor of $p(t)$ which contains the 8 zeros of $p(t)$ nearest to the value a, we chose

$$p_1(t) = (t - a)^8$$

as a starting approximation. Again, the monic polynomials $\psi_{n-l}^{(0)}(t)$, $1 \le l \le 8$, coincide with the normalized derivatives of $p(t)$ of appropriate order. We observe that usually all the zeros of $\hat{p}(t)$ are clustered in a small circle around the origin in the complex plane and numerical difficulties arise when the value of a approaches this circle which contains the remaining zeros of $p(t)$. For example, when $a = 1.2$ and $n = 200$ the spectral condition number of the resultant matrix generated by $s(t)$ and $p^\star(t)$ is of order 10^{11} on average. Table 3 reports the average number of iterations in the cases $a = 2$, $a = 1.5$, where we set $\eta = 10^{-2}$ and $itmax = 200$.

4. Conclusions. In conclusion, it seems that this is quite a promising approach. A representation of the multishift LR matrix iteration in a polynomial setting has been established. This provides the means for obtaining a globally convergent algorithm for the approximation of the coefficients of a factor of degree k of a polynomial of degree n. The complexity of our algorithm complemented with a suitable shift strategy is linear with respect to n; moreover, the storage needed is also linear with respect to n. In this way, from a computational point of view, the proposed algorithm compares favorably with more traditional approaches based on the reduction to the eigenvalue computation, especially when we consider polynomials of high degree. The computational results are clearly preliminary. In general, our experiments indicate that the algorithm has a good performance for moderate values of k. Moreover, the convergence behavior depends on the separation between the wanted and unwanted spectrum. Future research on this topic might include the use of our algorithm as a tool for treating the cluster case in the root-finding algorithms based on simultaneous approximations of the roots (the Durand–Kerner algorithm implemented in the CMLIB library and Aberth's method implemented by Bini [2]). In this respect a more rigorous criterion for choosing the value of η is also required. Investigations of the use of our algorithm in a multiprecision floating point arithmetic would also be interesting.

REFERENCES

[1] Z. BAI AND J. W. DEMMEL, *On a block implementation of the Hessenberg multishift QR iteration*, J. High-Speed Comput., 1 (1989), pp. 97–112.

[2] D. BINI, *Numerical computation of polynomial zeros by means of Aberth's method*, Numer. Algorithms, 13 (1996), pp. 179–200.

[3] D. BINI AND L. GEMIGNANI, *On the complexity of polynomial zeros*, SIAM J. Comput., 21 (1992), pp. 781–799.

[4] J. J. Buoni, P. A. Farrell, and A. Ruttan, *Parallel LU decomposition of upper Hessenberg matrices on a shared memory multiprocessor*, in Comput. Appl. Math., C. Brezinski and U. Kulish, eds., Elsevier, North–Holland, Amsterdam, 1992.

[5] C. Carstensen and T. Sakurai, *Simultaneous factorization of a polynomial by rational approximation*, J. Comput. Appl. Math., 61 (1995), pp. 165–178.

[6] T. J. Dekker and J. F. Traub, *An analysis of the shifted LR algorithm*, Numer. Math., 17 (1971), pp. 179–188.

[7] T. J. Dekker and J. F. Traub, *The shifted QR algorithm for Hermitian matrices*, Linear Algebra Appl., 4 (1971), pp. 137–154.

[8] A. A. Dubrulle and G. H. Golub, *A multishift QR iteration without computation of the shifts*, Numer. Algorithms, 7 (1994), pp. 173–181.

[9] T. L. Freeman and R. W. Brankin, *A divide and conquer method for polynomial zeros*, J. Comput. Appl. Math., 30 (1990), pp. 71–79.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1990.

[11] M. A. Jenkins and J. F. Traub, *A three-stage variable shift iteration for polynomial zeros and its relation to generalized Rayleigh iteration*, Numer. Math., 14 (1970), pp. 256–263.

[12] M. A. Jenkins and J. F. Traub, *A three-stage algorithm for real polynomials using quadratic iteration*, SIAM J. Numer. Anal., 7 (1970), pp. 545–566.

[13] J. Kautsky and G. H. Golub, *On the calculation of Jacobi matrices*, Linear Algebra Appl., 52/53 (1983), pp. 439–455.

[14] A. S. Householder, *The Numerical Treatment of a Single Nonlinear Equation*, McGraw–Hill, New York, 1970.

[15] H. Rutishauser, *Lectures on Numerical Mathematics*, Birkhäuser, Boston, MA, 1990.

[16] T. Sakurai, H. Sugiura, and T. Torri, *Numerical factorization of a polynomial by rational Hermite interpolation*, Numer. Algorithms, 3 (1992), pp. 411–418.

[17] A. Schönhage, *The Fundamental Theorem of Algebra in Terms of Computational Complexity*, Tech. report, Department of Mathematics, University of Tubingen, Federal Republic of Germany, 1982.

[18] G. W. Stewart, *On the companion operator for analytic functions*, Numer. Math., 18 (1971), pp. 26–43.

[19] D. S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.

[20] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 182–190, January 1998

012

EIGENVECTOR SLICING OF THE NONNEGATIVE MATRICES*

D. J. HARTFIEL†

Abstract. This paper shows how the set of nonnegative matrices can be seen as the union of smaller-dimensional sets called slices. For 2×2 matrices a model is drawn and it is shown how the model can be used to study the eigenvector behavior of a perturbed matrix.

Key words. nonnegative matrices, eigenvector model

AMS subject classifications. 15A48, 15A18, 15A12

PII. S0895479895298706

Let N be the set of $n \times n$ nonnegative matrices. If $A \in N$, by the Perron–Frobenius theory [5] A has at least one nonnegative eigenvalue ρ and a nonnegative left eigenvector $x \neq 0$ as well as a nonnegative right eigenvector $y \neq 0$. Thus,

$$xA = \rho x \quad \text{and} \quad Ay = \rho y.$$

A stochastic vector is a nonnegative vector with entries summing to one. Throughout this paper we assume, without loss of generality, that x and y are stochastic eigenvectors.

Of course, there can be many matrices which have x and y as their eigenvectors. Define a slice of N as

$$S(x, y) = \{A \in N\colon\ A \text{ has left eigenvector } x \text{ and right eigenvector } y\}.$$

Thus, it is clear that $N = \cup S(x, y)$, where the union is over all stochastic vectors x and y. So, N can be seen as a union of slices. A similar set is studied in [3], [12].

In this paper we give some geometrical description of a slice as well as how slices fit together to form N. A model is drawn for the 2×2 matrices and we show how this model gives a qualitative view of eigenvector behavior for a perturbed matrix. In studying problems posed in n-space it is often helpful to study small-dimensional cases for insight as well as for testing grounds for ideas and conjectures. See [10], [11] for related work.

Results. It is clear that N is a closed convex cone and its dimension [1], [2], or [4] is n^2. Any slice is also a closed convex cone. We determine its dimension below.

To this end, we define the intermediate convex set

$$S_1(x, y) = \{A \in S(x, y)\colon\ xA = x \quad \text{and} \quad Ay = y\}.$$

By simultaneous permutation of rows and columns of the matrices in $S_1(x, y)$ we can assume that

$$x = (x_1, x_2, 0, 0) \quad \text{and} \quad y = \begin{bmatrix} 0 \\ y_1 \\ y_2 \\ 0 \end{bmatrix},$$

*Received by the editors September 12, 1995; accepted for publication (in revised form) by V. Mehrmann January 31, 1997.
http://www.siam.org/journals/simax/19-1/29870.html
†Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (hartfiel@math.tamu.edu).

where

(i) each of x and y are partitioned into subvectors having n_1, n_2, n_3, n_4 components. In this description, zero components are allowed.

(ii) No entry in x_1, x_2, y_1, y_2 is zero.

If we partition the matrices in $S_1(x, y)$ compatibly to the vectors and observe the equations

$$\begin{aligned} &xA = x, \\ &Ay = y \quad \text{for any} \quad A \in S_1(x, y), \end{aligned}$$

we note that

$$A = \begin{bmatrix} B & 0 & 0 & 0 \\ C & D & 0 & 0 \\ X & F & G & W \\ Y & 0 & 0 & Z \end{bmatrix},$$

where the main diagonal blocks are $n_1 \times n_1$, $n_2 \times n_2$, $n_3 \times n_3$, $n_4 \times n_4$, respectively. Of course, some of these blocks may not appear.

The equations in the subblocks of A must satisfy

$$\begin{aligned} &(1) \quad (x_1, x_2) \begin{bmatrix} B \\ C \end{bmatrix} = x_1, \\ &(2) \quad x_2 D = x_2, \\ &(3) \quad D y_1 = y_1, \\ &(4) \quad [F\ G] \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = y_2. \end{aligned}$$

Conversely, any matrices which satisfy (1) through (4) can be used to construct, in the obvious way, a matrix $A \in S_1(x, y)$.

Define

$$\begin{aligned} &(1') \quad \mathcal{B} = \{M\colon\ M \text{ is an } (n_1 + n_2) \times n_1 \text{ matrix}\}, \\ &(2') \quad \mathcal{D} = \{M\colon\ M \text{ is an } n_2 \times n_2 \text{ matrix}\}, \\ &(3') \quad \mathcal{E} = \{M\colon\ M \text{ is an } n_3 \times (n_2 + n_3) \text{ matrix}\}, \\ &(4') \quad \mathcal{X} = \{M\colon\ M \text{ is an } (n_3 + n_4) \times (n_1 + n_4) \text{ matrix}\}, \end{aligned}$$

and, finally,

$$\begin{aligned} \mathcal{A} = \{&M\colon\ M \text{ is an } n \times n \text{ matrix having the same zero submatrices} \\ &\text{as that of the partitioned matrix } A\}. \end{aligned}$$

Note that all these sets are vector spaces and thus so is the direct sum $B \oplus \mathcal{D} \oplus \mathcal{E} \oplus \mathcal{X}$. Define

$$\begin{aligned} &\varphi\colon\ \mathcal{A} \longrightarrow \mathcal{B} \oplus \mathcal{D} \oplus \mathcal{E} \oplus \mathcal{X} \quad \text{by} \\ &\varphi(A) = \left(\begin{bmatrix} B \\ C \end{bmatrix}, \quad D, \quad [F, G], \quad \begin{bmatrix} X & W \\ Y & Z \end{bmatrix} \right), \end{aligned}$$

where A is partitioned as before,

$$A = \begin{bmatrix} B & 0 & 0 & 0 \\ C & D & 0 & 0 \\ X & F & G & W \\ Y & 0 & 0 & Z \end{bmatrix}.$$

It is easy to see that φ: $\mathcal{A} \longrightarrow \mathcal{B} \oplus \mathcal{D} \oplus \mathcal{E} \oplus \mathcal{X}$ is an isomorphism and this isomorphism preserves convex sets and dimension. Thus, we can compute $\dim S_1(x,y)$ by computing $\dim \varphi(S_1(x,y))$, which we do. In our formula we use the notation

$$\begin{aligned}
\mathcal{B}_+ &= \{M\colon\ M \text{ is nonnegative and satisfies equation (1)}\},\\
\mathcal{D}_+ &= \{M\colon\ M \text{ is nonnegative and satisfies both equations (2) and (3)}\},\\
\mathcal{E}_+ &= \{M\colon\ M \text{ is nonnegative and satisfies equation (4)}\},\\
\mathcal{X}_+ &= \{M\colon\ M \text{ is nonnegative and in } \mathcal{X}\}.
\end{aligned}$$

THEOREM 1. $\dim\ S_1(x,y) = \dim \mathcal{B}_+ + \dim \mathcal{D}_+ + \dim \mathcal{E}_+ + \dim \mathcal{X}_+$.

Proof. Note that

$$\varphi(S_1(x,y)) = \mathcal{B}_+ \oplus \mathcal{D}_+ \oplus \mathcal{E}_+ \oplus K_+.$$

It is known that the dimension of a direct sum is the sum of the individual dimensions. The result follows. □

We now give formulas for the dimensions of $\mathcal{B}_+$, $\mathcal{D}_+$, $\mathcal{E}_+$, and $\mathcal{X}_+$.

THEOREM 2. *Assuming the sets below are nonempty,* $\dim \mathcal{B}_+ = n_1(n_1+n_2-1)$, $\dim \mathcal{D}_+ = (n_2-1)(n_2-1)$, $\dim \mathcal{E}_+ = n_3(n_2+n_3-1)$, *and* $\dim \mathcal{X}_+ = (n_3+n_4)(n_1+n_4)$.

Proof. We need only compute the dimensions of each of the corresponding sets of nonnegative matrices. In doing this, we use that if $z = (z_i)$ is a vector, diag z is the diagonal matrix with main diagonal entries $z_1, \ldots, z_n$, respectively.

Case for $\mathcal{B}_+$ and $\mathcal{E}_+$: Define $X = \operatorname{diag}(x_1, x_2)$ and $X_1 = \operatorname{diag} x_1$. A straightforward calculation shows that $\begin{bmatrix} B \\ C \end{bmatrix}$ satisfies (1) if and only if $X \begin{bmatrix} B \\ C \end{bmatrix} X_1^{-1}$ has all its column sums equal to one. Now, as shown in [7], the set of $(n_1+n_2)\times n_1$ nonnegative matrices with all column sums one has dimension $n_1(n_1+n_2-1)$. Since the map $\theta(M) = XMX_1^{-1}$ is an isomorphism, $\dim \mathcal{B}_+ = n_1(n_1+n_2-1)$.

The same argument can be used to prove that $\dim \mathcal{E}_+ = n_3(n_2+n_3-1)$.

Case for $\mathcal{D}_+$: Define $X_2 = \operatorname{diag} x_2$ and $Y_1 = \operatorname{diag} y_1$. A direct calculation shows that $D \in \mathcal{D}_+$ if and only if D is nonnegative and $X_2 D Y_1$ has row sum and column sum vectors $x_2 Y_1$ and $X_2 y_1$, respectively. In [7], it is shown that the dimension of the set of nonnegative matrices, with row sum and column sum vectors $x_2 Y_1$ and $X_2 y_1$, respectively, is $(n_2-1)(n_2-1)$. Since the map $\theta(M) = X_2 M Y_1$ is an isomorphism, $\dim \mathcal{D}_+ = (n_2-1)(n_2-1)$.

Case for $\mathcal{X}_+$: It is clear that this set has dimension $(n_3+n_4)(n_1+n_4)$.

Putting these together yields the result. □

To simplify some of the computations which follow, we now introduce a result about convex sets. This result requires a few technical descriptions.

Let $\mathcal{M}$ be the set of $r \times s$ matrices. A convex set F of nonnegative matrices in $\mathcal{M}$ is called a *flat* if there is a linear functional φ on $\mathcal{M}$ and a nonzero constant c such that $\varphi(A) = c$ for all $A \in F$.

For any convex set of nonnegative matrices K, $K \neq \{0\}$, define

$$K^c = \cup \rho K, \quad \text{where} \quad \rho \geq 0$$

and $\rho K = \{\rho k\colon\ k \in K\}$. It is easily seen that K^c is a convex cone.

Vectors $w_0, \ldots, w_n$ are independent provided that $w_1 - w_0, \ldots, w_n - w_0$ are linearly independent (equivalently, we could have subtracted w_i for any i). It is known that convex $\{w_0, \ldots, w_n\}$, called a simplex, is a convex polytope with vertices $w_0, \ldots, w_n$.

The dimension of this simplex is n. Further, the dimension of a convex set C is precisely the dimension of the largest, in dimension, simplex which is a subset of C.

The following result is intuitively clear.

LEMMA 1. *Let F be a nonzero flat. Then* $\dim F^c = \dim F + 1$.

Proof. Suppose $\dim F = d$. Then F has $d+1$ independent vectors, say $A_0, \ldots, A_d$. We show, arguing by contradiction, that these vectors are linearly independent.

Suppose there are scalars, not all zero, satisfying

$$\alpha_0 A_0 + \cdots + \alpha_d A_d = 0.$$

Without loss of generality, suppose $\alpha_0 \neq 0$. Solving for A_0 yields

$$A_0 = \beta_1 A_1 + \cdots + \beta_d A_d,$$

where $\beta_i = -\alpha_i/\alpha_0$. Since F is a flat, there is a linear functional and a nonzero constant c such that $\varphi(A) = c$ for all $A \in F$. Thus, computing φ of both sides of the equation yields

$$\begin{aligned} \varphi(A_0) &= \beta_1 \varphi(A_1) + \cdots + \beta_d \varphi(A_d) \quad \text{or} \\ c &= \beta_1 c + \cdots + \beta_d c. \end{aligned}$$

Hence, $\beta_1 + \cdots + \beta_d = 1$. But now we can write the initial equation involving the β_i's as

$$0 = \beta_1(A_1 - A_0) + \cdots + \beta_d(A_d - A_0)$$

which implies that $A_0, A_1, \ldots, A_d$ is not an independent set of vectors. This is a contradiction from which it follows that $A_0, A_1, \ldots, A_d$ is a linearly independent set.

Now, using that $A_0, A_1, \ldots, A_d$ is a linearly independent set it follows that 0, $A_0, \ldots, A_d$ is an independent set in F^c. Thus, $\dim F^c \geq d + 1$. We will show that equality holds.

Consider $W = \operatorname{span}\{A_0, A_1, \ldots, A_d\}$. If $F^c \not\subseteq W$, then there is a vector $A_{d+1} \in F^c$ such that $A_{d+1} \notin W$. By scaling, we can assume that $A_{d+1} \in F$. But then $A_0, A_1, \ldots, A_{d+1}$ is a set of linearly independent vectors in F and thus $\dim F > d$, which is a contradiction. Thus, $F^c \subseteq W$. Since $\dim W = d + 1$ it then follows that $\dim F^c = d + 1$. □

We now apply this result to compute the dimensions of several convex sets.

COROLLARY 1. $\dim S(x, y) \equiv \dim S_1(x, y) + 1$.

Proof. Define φ on the set of $n \times n$ matrices by $\varphi(A) = xAe$, where e is the vector all of whose entries are one. Then φ is a linear functional. Further, if $A \in S_1(x, y)$, then $\varphi(A) = xAe = xe = 1$. Thus, $S_1(x, y)$ is a flat.

Now, noting that $S_1(x, y)^c = S(x, y)$, the result follows from the theorem. □

Define

$$S_{\#}(x, y) = \left\{ A \in T\colon\ \|A\| = \sum_{i,j} a_{ij} = 1 \right\}.$$

It is clear that $S_{\#}(x, y)$ is a convex set. (In general, a subscript # in set notation will indicate that subset whose matrices have entries summing to one.)

COROLLARY 2. $\dim S_{\#}(x, y) = \dim S_1(x, y)$.

Proof. Define φ on the set of $n \times n$ matrices by $\varphi(A) = \sum_{i,j} a_{ij}$. This is a linear functional. And $\varphi(A) = 1$ for all $A \in S_{\#}(x,y)$, so $S_{\#}(x,y)$ is a flat. The corollary follows by noting that $S_{\#}(x,y)^c = S(x,y)$. □

This completes our work on individual slices. To see how these slices fit together to form N, we need to introduce a connecting set. Define the set of rank one matrices

$$R = \{A\colon A = yx, \text{ where } x, y \text{ are } 1 \times n \text{ and } n \times 1 \text{ stochastic vectors, respectively.}\}$$

We show R intersects each $S(x,y)$.

THEOREM 3. *The set R intersects $S(x,y)$ at $A = yx$.*

Proof. Note that $xA = x(yx) = (xy)x$ and that $Ay = (yx)y = y(xy)$. Since xy is a scalar, $A \in S(x,y)$. Thus, R intersects $S(x,y)$ at A. □

We now give a description of R.

THEOREM 4. *R is convex in x and convex in y.*

Proof. We show that R is convex in y. For this let $y, \bar{y}$ be $n \times 1$ stochastic vectors and x be a $1 \times n$ stochastic vector. Then $[\alpha y + (1-\alpha)\bar{y}]x = \alpha yx + (1-\alpha)\bar{y}x$. Thus, R is convex in y. □

In the next theorem we compute the manifold dimension of R.

THEOREM 5. *R is a manifold, with boundary, of dimension $2(n-1)$.*

Proof. Let $\alpha_1, \ldots, \alpha_{n-1}, \beta_1, \ldots, \beta_{n-1}$ be nonnegative variables. Set $\alpha = \alpha_1 + \cdots + \alpha_{n-1}$ and $\beta = \beta_1 + \cdots + \beta_{n-1}$. Define

$$H = \{(\alpha_1, \ldots, \alpha_{n-1}, \beta_1, \ldots, \beta_{n-1})\colon \alpha \le 1 \text{ and } \beta \le 1\} \subseteq R^{2(n-1)}.$$

Then H is a manifold, with boundary, such that its dimension is $2(n-1)$. The map

$$f(\alpha_1, \ldots, \alpha_{n-1}, \beta_1, \ldots, \beta_{n-1}) = \begin{bmatrix} \alpha_1\beta_1 & \alpha_1\beta_2 & \ldots & \alpha_1\beta_{n-1} & \alpha_1(1-\beta) \\ \alpha_2\beta_1 & \alpha_2\beta_2 & \ldots & \alpha_2\beta_{n-1} & \alpha_2(1-\beta) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (1-\alpha)\beta_1 & (1-\alpha)\beta_1 & \ldots & (1-\alpha)\beta_{n-1} & (1-\alpha)(1-\beta) \end{bmatrix}$$

shows that R is a manifold as described in the lemma. □

We now describe how N is put together using slices and R.

We know N is the union of slices, each slice being a closed convex cone. The dimensions of the slices can vary; however, the dimensions of slices corresponding to positive eigenvectors are the same dimension $(n-1)^2 + 1$. These slices are precisely those that intersect the interior of N. Further, by the Perron–Frobenius theory, nonnegative matrices with positive stochastic eigenvectors cannot have additional nonnegative stochastic eigenvectors. Thus, slices which intersect the interior of N can not intersect each other. Yet, each $S(x,y)$ intersects R at yx and to some extent this shows how the slices lie together. As given below, if the intersections of two slices with R are close, so are the corresponding eigenvectors, and vice versa. To show this, we use the ℓ_1-norm and the ℓ_∞-norm. Recall

$$\|A\|_1 = \max_{w \neq 0} \frac{\|wA\|_1}{\|w\|_1}, \quad \|A\|_\infty = \max_{w \neq 0} \frac{\|wA\|_\infty}{\|w\|_\infty}.$$

And $\|A\|_1 = \max_i \sum_k |a_{ik}|$, $\|A\|_\infty = \max_j \sum_k |a_{kj}|$.

THEOREM 6. *Let $x, \bar{x}$ be $n \times 1$ and $y, \bar{y}$ be $1 \times n$ stochastic vectors, respectively.*

(i) *If $\max\{\|xy - \bar{x}\,\bar{y}\|_\infty, \|xy - \bar{x}\,\bar{y}\|_1\} = \epsilon$, then $\|y - \bar{y}\|_\infty \le \epsilon$ and $\|x - \bar{x}\|_\infty \le \epsilon$.*

TABLE 1
Comparisons of dimensions of various convex sets.

n	$N\#$	R	$S\#(x,y)$
2	3	2	1
3	8	4	4
4	15	6	9
5	24	8	16

(ii) *If* $\|x-\bar{x}\|_1 \le \epsilon$ *and* $\|y-\bar{y}\|_1 \le \delta$, *then* $\|yx-\bar{y}\,\bar{x}\|_1 \le \epsilon+\delta$.

Proof. For (i),

$$\epsilon \ge \|xy-\bar{x}\,\bar{y}\|_\infty \ge \|w(xy-\bar{x}\,\bar{y})\|_\infty, \quad \text{where}$$

$w=(1,1,\dots,1)$. Thus, since $wx=w\bar{x}=1$, $\epsilon \ge \|y-\bar{y}\|_\infty$.

Further,

$$\begin{aligned}\epsilon \ge \|xy-\bar{x}\,\bar{y}\|_1 &= \|(xy)^t-(\bar{x}\,\bar{y})^t\|_\infty\\ &= \|y^tx^t-\bar{y}^t\bar{x}^t\|_\infty\\ &\ge \|x^t-\bar{x}^t\|_\infty \quad \text{as shown previously.}\end{aligned}$$

So, $\epsilon \ge \|x-\bar{x}\|_\infty$.

For (ii), set $\bar{x}-x=\varepsilon$ and $\bar{y}-y=d$. Then

$$\begin{aligned}\|yx-\bar{y}\,\bar{x}\|_1 &= \|yx-(y+d)(x+\varepsilon)\|_1\\ &= \|y\varepsilon+d\bar{x}\|_1\\ &\le \|y\varepsilon\|_1+\|d\bar{x}\|_1\\ &\le \|\varepsilon\|_1+\|d\|_1\\ &\le \epsilon+\delta. \qquad \square\end{aligned}$$

To obtain a model of the description of N in terms of slices, it is best to look at

$$N_\# = \{A \in N\colon\ \|A\|=1\}.$$

This is a convex set and its dimension can be calculated, as in the corollaries, as n^2-1.

Before proceeding, it is helpful to look at Table 1 to compare dimensions of the various convex sets.

By examining the table it is clear that we should be able to provide a sketch of how the slices fit in $N_\#$ when $n=2$. Here $N_\# = \cup S_\#(x,y)$, the union over all stochastic vectors x and y.

The set $N_\#$ is a 3-simplex with vertices diagrammed in Fig. 1.

The manifold R is two-dimensional and can be viewed as those points obtained by moving the segment $\overline{AB}$ to the segment $\overline{CD}$, keeping the endpoints at the same distance from A to D as from B to C. (So we have a twisted sheet.)

We now put in the slices $S_\#(x,y)$, over all stochastic vectors x and y. Note that $E=\begin{bmatrix}1/2 & 0\\ 0 & 1/2\end{bmatrix}$ is in all of the slices. We will call it the anchor. The matrix $yx \in S_\#(x,y)$, and since $\dim S_\#(x,y)=1$, $S_\#(x,y)$ is the line segment in $N_\#$ beginning at E and passing through yx (see Fig. 2).

To use this model in a qualitative study of perturbation, choose $P \in N_\#$. Then $P \in S_\#(x,y)$ for some stochastic eigenvectors x and y of P. Thus, P is on the segment

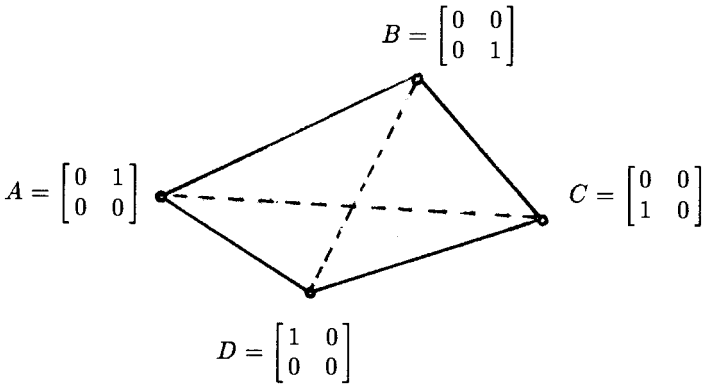


FIG. 1. *A drawing of* $N_{\#}$.

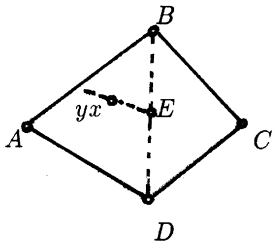


FIG. 2. $S_{\#}(x, y)$ *placed in* $N_{\#}$. *The manifold* R *is the twisted sheet determined by the solid lines.*

in $N_{\#}$ determined by E and yx. Perturb P to $\overline{P} \in N_{\#}$. Then $\overline{P} \in S_{\#}(\bar{x}, \bar{y})$ is on the segment in $N_{\#}$ containing E and $\bar{y}\,\bar{x}$. Now to see the effect of this perturbation we see how far $\bar{y}\,\bar{x}$ is from yx. (If $\bar{y}\,\bar{x}$ is close to yx, then $\bar{y}$ is close to y and $\bar{x}$ is close to x.)

We now provide some examples showing how to use $N_{\#}$ and slices to study perturbation of eigenvectors in a qualitative way.

Example. For $n = 2$, let P be on the opposite side of R from the anchor E, e.g., $P = \left[\begin{smallmatrix} 0 & 1/2 \\ 1/2 & 0 \end{smallmatrix}\right]$. Suppose $P \in S_{\#}(x, y)$.

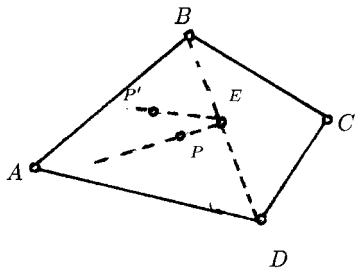


FIG. 3. *View of sensitive areas in* $N_{\#}$.

Then perturbing P to $P' \in S_{\#}(x', y')$ shows that $y'x'$ is not perturbed from yx as much as if P and P' were on the same side of R as the anchor E, e.g., if $P = \left[\begin{smallmatrix} 1/2 - \epsilon & \epsilon \\ \epsilon & 1/2 - \epsilon \end{smallmatrix}\right]$ for small ϵ (see Fig. 3).

Example. An interesting observation can be made for $n = 2$. Let $P = \begin{bmatrix} \frac{1}{2} - \epsilon & 2\epsilon \\ 0 & \frac{1}{2} - \epsilon \end{bmatrix}$. Note that P has eigenvalues $\lambda_1 = \lambda_2 = \frac{1}{2} - \epsilon$. Viewing the segment in $N_\#$ determined by P and E and its intersection yx with R, then perturbing that segment shows that when ϵ is close to zero, the corresponding eigenvectors are sensitive while if ϵ is close to $\frac{1}{2}$, the corresponding eigenvectors are not that sensitive. Yet, the perturbed matrix has eigenvalues which can be very close. Recall that close eigenvalues can indicate sensitive eigenvectors (see Fig. 4).

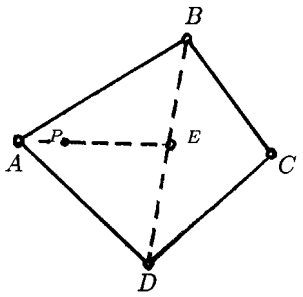


FIG. 4. *View of perturbation of matrices with close eigenvalues.*

Example. For an example with n larger, we will fix $x = (1, \ldots, 1)^t$ and set $N(x) = \{A \in N\colon A \text{ has right eigenvector } x\}$ and $R(x) = N(x) \cap R$ which, using Theorem 4, is convex. Note that if $A \in N_\#(x)$, then A has all its row sums equal to $\frac{1}{n}$. Thus, using well-known results about stochastic matrices, $\dim N_\#(x) = n(n-1)$.

For $n = 2$, $N(x)$ and $R(x)$ are shown below (see Fig. 5). A similar view is given in [6].

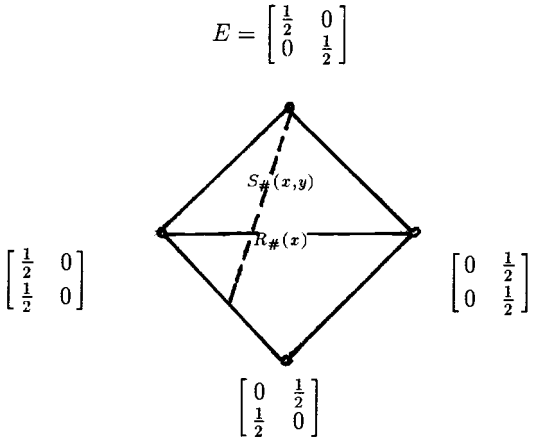


FIG. 5. *Sketch of* $N_\#(x)$.

In [8] it is shown that if $A \in S_\#(x, y)$, then y is sensitive to changes in A if and only if A has a subdominant eigenvalue close to $\frac{1}{n}$. And, in [9] it is shown that if $A \in S_\#(x, y)$ has a subdominant eigenvalue close to $\frac{1}{n}$, then A nearly has two isolated

main diagonal submatrices, e.g.,

$$A = \begin{bmatrix} A_{11} & \mathcal{E}_{12} & \mathcal{E}_{13} \\ \mathcal{E}_{21} & A_{22} & \mathcal{E}_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix},$$

where the $\mathcal{E}_{ij}$ are small. Thus, y is sensitive precisely when A nearly has two isolated main diagonal submatrices.

To see this phenomenon we will observe a simpler case below (see Fig. 6).

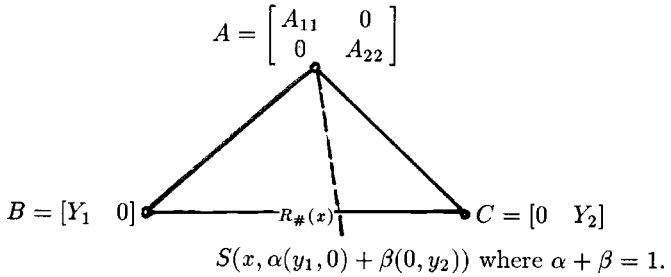


FIG. 6. *Partial sketch of* $N_{\#}(x)$.

Here Y_1 has rows $\frac{1}{n}(y_1, 0)$ where $y_1 A_{11} = \frac{1}{n} y_1$, Y_2 has rows $\frac{1}{n}(0, y_2)$ where $y_2 A_{22} = \frac{1}{n} y_2$, and y_1, y_2 are both stochastic vectors. Note that $A, B \in S_{\#}(x, (y_1, 0))$ and $A, C \in S_{\#}(x, (0, y_2))$. From this we see that if small changes are made in A, staying in the convex set determined by A, B, and C, then there are large changes in the corresponding left eigenvectors, as described by the results in [8] and [9].

In conclusion, drawings that provide a qualitative view are important in areas where everything is given quantitatively. The work in this paper does some of that in the study of matrix perturbation on eigenvectors. In addition, this paper can provide an outline for techniques in developing other such drawings.

Additional material on this subject can be found in the references.

REFERENCES

[1] A. BERMAN, *Cones, Matrices and Mathematical Programming*, Springer-Verlag, New York, 1973.
[2] A. BRONDSTED, *An Introduction to Convex Polytopes*, Springer-Verlag, New York, 1982.
[3] R. A. BRUALDI AND H. J. RYSER, *Combinatorial Matrix Theory*, Cambridge University Press, New York, 1991.
[4] H. G. EGGLESTON, *Convexity*, Cambridge University Press, New York, 1958.
[5] F. R. GANTMACHER, *The Theory of Matrices*, Volume 2, Chelsea, New York, 1964.
[6] R. E. GONZALEZ, *A Geometric Study of Certain Stochastic Semigroups*, Dissertation, University of Houston, Houston, TX, 1983.
[7] D. J. HARTFIEL, *A study of convex sets of stochastic matrices induced by probability vectors*, Pacific J. Math., 52 (1974), pp. 405–418.
[8] C. D. MEYER, *Sensitivity of the stationary distribution of a Markov chain*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 715–728.
[9] C. D. MEYER AND D. J. HARTFIEL, *On the structure of stochastic matrices with a subdominant eigenvalue near* 1, Linear Algebra Appl., to appear.
[10] N. J. PULLMAN, *A geometric approach to the theory of nonnegative matrices*, Linear Algebra Appl., 4 (1971), pp. 297–312.
[11] N. J. PULLMAN, *The geometry of finite Markov chains*, Canad. Math. Bull., 8 (1965), pp. 345–357.
[12] H. J. RYSER, *Combinatorial Mathematics*, American Mathematical Society, Providence, RI, 1963.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 191–204, January 1998

013

EUCLIDEAN NORM MINIMIZATION OF THE SOR OPERATORS*

APOSTOLOS HADJIDIMOS† AND MICHAEL NEUMANN‡

Abstract. Because the spectral radius is only an asymptotic measure of the rate of convergence of a linear iterative method, Golub and dePillis [*Toward an effective two-parameter method,* in Iterative Methods for Large Linear Systems, Academic Press, New York, 1990] have raised in a recent paper the question of determining, for each $k \geq 1$, a relaxation parameter $\omega \in (0, 2)$ and a pair of relaxation parameters ω_1 and ω_2 which minimize the Euclidean norm of the kth power of the SOR and MSOR iteration matrices, respectively, associated with a real symmetric positive definite matrix with "Property A." Here we use a reduction of these operators which they derived from the SVD of the associated block Jacobi matrix to obtain the minimizing relaxation parameters for the case $k = 1$ for both operators. We conclude the paper with two brief sections in which we assess what our results imply.

Key words. linear systems, iterative methods, relaxation methods, optimal convergence

AMS subject classification. 65F10

PII. S0895479896300498

1. Introduction and preliminaries. In a relatively recent paper, Golub and dePillis [1] raise, in light of new reductions of the SOR and MSOR iteration matrices, the recurring question of minimizing the Euclidean norm of the kth power of the SOR and MSOR operators as a function of the relaxation parameter(s). This question is of interest because for small values of k, it is the norm of the kth power of the iteration matrix which governs the rate of convergence in the initial stages of the iteration rather than the spectral radius of the iteration matrix which is an asymptotic measure.

The new reductions of the SOR and MSOR iteration operators which Golub and dePillis carry out are achieved using the SVD (see, e.g., [2]). It is based on an idea of Lanczos [4] who used SVD to reduce a real symmetric positive definite matrix possessing "Property A." Golub and dePillis derive explicit expressions for both the block SOR and the block MSOR operators associated with the 2×2 block partitioning of A denoted by $\mathcal{L}_\omega$ and $\mathcal{L}_{\omega_1,\omega_2}$, respectively. As a by-product of their analysis they also derive Young's famous relationship (see, e.g., [7], [6], and [10])

$$(\lambda + \omega - 1)^2 \;=\; \omega^2\mu^2\lambda \tag{1}$$

connecting the eigenvalues μ and λ of the block Jacobi operator B and of the block SOR operator $\mathcal{L}_\omega$ associated with A, and also the more general relationship (see, e.g., [8] and [10])

$$(\lambda + \omega_1 - 1)(\lambda + \omega_2 - 1) = \omega_1\omega_2\mu^2\lambda \tag{2}$$

relating the eigenvalues μ and λ of the block Jacobi operator B and the block MSOR operator $\mathcal{L}_{\omega_1,\omega_2}$, respectively. In (1) and (2), ω, and ω_1 and ω_2 are the relaxation

* Received by the editors February 15, 1996; accepted for publication (in revised form) by M. Eiermann February 3, 1997.
http://www.siam.org/journals/simax/19-1/30049.html

† Department of Computer Sciences, Purdue University, West Lafayette, IN 47907 (hadjidim@cs.purdue.edu). The research of this author was supported by NSF grant CCR 86–19817, AFOSR grant 91–F49620, and ARPA grant DAAH04–94–G–0010.

‡ Department of Mathematics, University of Connecticut, Storrs, CT 06269–3009 (neumann@math.uconn.edu). The research of this author was supported in part by NSF grant DMS–9007030.

parameters associated with the SOR and MSOR methods, respectively. In this work we adopt much of the notation used in [1]. Let

$$A = \begin{bmatrix} I_p & -M \\ -M^T & I_q \end{bmatrix} =: I - B \in \mathbb{R}^{n,n}, \tag{3}$$

where $M \in \mathbb{R}^{p,q}$ with $p + q = n$ and $p \geq q$. Suppose that

$$M = U\Sigma V \tag{4}$$

is the SVD of M, where $U \in \mathbb{R}^{p,p}$ and $V \in \mathbb{R}^{q,q}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{p,q}$ is the (diagonal) matrix of singular values s_i, $i = 1, \ldots, q$, with $s_1 \geq s_2 \geq \cdots \geq s_q \geq 0$, which has the form

$$\Sigma = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & & s_q \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{5}$$

Then Jacobi, SOR, and MSOR operators associated with the block partitioning of A in (3) are the matrix B defined via (3), the matrix

$$\mathcal{L}_\omega = \begin{bmatrix} (1-\omega)I_p & \omega M \\ \omega(1-\omega)M^T & (1-\omega)I_q + \omega^2 M^T M \end{bmatrix}, \tag{6}$$

and the matrix

$$\mathcal{L}_{\omega_1,\omega_2} = \begin{bmatrix} (1-\omega_1)I_p & \omega_1 M \\ \omega_2(1-\omega_1)M^T & (1-\omega_2)I_q + \omega_1\omega_2 M^T M \end{bmatrix}, \tag{7}$$

respectively. Golub and dePillis apply the SVD factorization of M given in (4), with Σ in (5), to obtain

$$\mathcal{L}_\omega = QP^T\Delta(\omega)PQ^T \quad \text{and} \quad \mathcal{L}_{\omega_1,\omega_2} = QP^T\Delta(\omega_1,\omega_2)PQ^T, \tag{8}$$

where $Q = \operatorname{diag}(U, V)$ and P is an appropriate permutation matrix. The matrices $\Delta(\omega)$ and $\Delta(\omega_1, \omega_2)$ in (8) have the block diagonal forms

$$\Delta(\omega) = \begin{bmatrix} \Delta_1(\omega) & & & \\ & \ddots & & \\ & & \Delta_q(\omega) & \\ & & & (1-\omega)I_{p-q} \end{bmatrix}, \tag{9}$$

$$\Delta(\omega_1,\omega_2) = \begin{bmatrix} \Delta_1(\omega_1,\omega_2) & & & \\ & \ddots & & \\ & & \Delta_q(\omega_1,\omega_2) & \\ & & & (1-\omega_1)I_{p-q} \end{bmatrix},$$

where

$$\Delta_i(\omega) = \begin{bmatrix} 1-\omega & \omega s_i \\ \omega(1-\omega)s_i & 1-\omega+\omega^2 s_i^2 \end{bmatrix}, \tag{10}$$

$$\Delta_i(\omega_1,\omega_2) = \begin{bmatrix} 1-\omega_1 & \omega_1 s_i \\ \omega_2(1-\omega_1)s_i & 1-\omega_2+\omega_1\omega_2 s_i^2 \end{bmatrix}, \quad i = 1,\dots,q.$$

Note. If $q \geq p$, then the roles of p and q in (9) and (10) and also that of ω_1 and ω_2 in the last diagonal blocks of (9) are interchanged.

In view of (6)–(10), the questions from [1] cited at the beginning of this paper can be recast as follows:

Problem I. Determine

$$\begin{aligned} \min_{\omega\in(0,2)} \| \mathcal{L}_\omega^k \|_2 &= \min_{\omega\in(0,2)} \| \Delta^k(\omega) \|_2 \\ &= \min_{\omega\in(0,2)} \left\{ \max \left\{ \max_{1\leq i\leq \min\{p,q\}} \| \Delta_i^k(\omega) \|_2, |1-\omega|^k \right\} \right\}. \end{aligned} \tag{11}$$

Problem II. Determine

$$\begin{aligned} \min_{\omega_1,\omega_2(0,2)} \| \mathcal{L}_{\omega_1,\omega_2}^k \|_2 &= \min_{\omega_1,\omega_2\in(0,2)} \| \Delta^k(\omega_1,\omega_2) \|_2 \\ &= \min_{\omega_1,\omega_2\in(0,2)} \left\{ \max \left\{ \max_{1\leq i\leq \min\{p,q\}} \| \Delta_i^k(\omega_1,\omega_2) \|_2, |1-\omega_1|^k \text{ or } |1-\omega_2|^k \right\} \right\}. \end{aligned} \tag{12}$$

The restrictions on the relaxation factors ω, ω_1, and ω_2 to the interval (0,2) come, of course, from the necessary and sufficient conditions for the powers of the iteration matrices $\mathcal{L}_\omega$ and $\mathcal{L}_{\omega_1,\omega_2}$ associated with the 2-cyclic consistently ordered and real symmetric positive definite matrix A in (3) to converge (see Theorems 6.2.2 and 8.3.2 in Young [10]) to the zero matrix. Also, only one of the terms $|1-\omega_1|^k$, $|1-\omega_2|^k$ is needed in (12), depending on whether $p \geq q$ or $q \geq p$.

In this paper we completely settle Problems I and II in the case $k = 1$. For this case Young [8], [9], [10], Young and Kincaid [11], and Young, Wheeler, and Downing [12] have found some very interesting initial results and observations, some of which, although analyzed and studied in Young's book [10], have gaps in their proofs or are only based on numerical evidence. Thus, motivated by the work of Golub and dePillis [1] and using as a guide the analysis in the works by Young and Young and coauthors, we have sought to generalize and extend the existing results further and also to fill in the gaps in the analysis. This we do in section 2 for the SOR operator and in section 3 for the MSOR operator.

In section 4 we provide lower bounds for the minima of Problems I and II in terms of the solution to these problems when, instead of the minimization in the Euclidean norm, the minimization is done with respect to the energy norm. Actually, these bounds can be found in Young [10] but seem relatively unknown, and they have a somewhat negative implication concerning the minimum established for Problem I, namely, that the minimum is bounded *below* by the spectral radius $\rho(B)$ of the Jacobi iteration matrix. This is *not so* with respect to the minimum established for Problem II, namely, that this minimum is bounded *above* by $\rho(B)$.

Finally, in section 5 we compare a numerical example that is given in [1], where $\| \mathcal{L}_{\omega_1,\omega_2}^{50} \|_2$ was minimized computationally, with the minimum of $\| \mathcal{L}_{\omega_1,\omega_2} \|_2^{50}$, where $\| \mathcal{L}_{\omega_1,\omega_2} \|_2$ is minimized using the results of this paper. The comparison shows that the former value is only very slightly better than the latter one. This may indicate

that, at least in cases of practical interest and when the spectral radius of the Jacobi iteration matrix is close to one, in order to save unnecessary calculations it is better to use directly the theoretical results of this paper regarding the value of the minimum of $\| \mathcal{L}_{\omega_1,\omega_2} \|_2$ rather than try and determine the minimum value of $\| \mathcal{L}^k_{\omega_1,\omega_2} \|_2$ for large k. Actually, our theoretical results indicate that performing the first few iterations with the iteration operator whose Euclidean norm has been minimized as a function of the relaxation parameter(s), rather than its spectral radius, is only beneficial if the MSOR operator is to be used and when the spectral radius of the Jacobi operator is close to one. Thus, for example, accelerating the initial iteration using the SOR iteration matrix whose Euclidean norm is minimal does not seem to yield much benefit over the use of the SOR iteration operator whose spectral radius is optimal.

2. Euclidean norm minimization of the SOR operator. For $k = 1$, Problem I in (11) was studied by Young on pages 245–247 in [10]. The result relevant to the work in this section is as follows: *Under the assumptions made and the notation introduced so far,* $\| \mathcal{L}_\omega \|_2 < 1$ *if and only if* $\omega < \frac{2(1-t)^{1/2}}{t+(1-t)^{1/2}}$, *where* $t := \rho^2(B)$.

In this section we solve completely Problem I as stated in (11) for the case $k = 1$. Our solution requires Lemma 2.1 and Theorem 2.2. The proof of Lemma 2.1 is relatively easy, while the proof of the main Theorem 2.2 is based, among others, on a series of intermediate results that can be found in [5] and to which the interested reader is referred.

LEMMA 2.1. *Let* $A \in \mathbb{R}^{n,n}$ *be a symmetric positive definite matrix having "Property A" and the block form* (3). *Then, Problem* I *for* k=1 *is equivalent to the determination of the quantity*

$$\widehat{\delta}^2 := \delta^2(\widehat{\omega}) \equiv \min_{\omega\in(0,2)} \delta^2 \equiv \min_{\omega\in(0,2)} \frac{1}{2}\left[T(t) + \left[T^2(t) - 4C\right]^{\frac{1}{2}}\right], \tag{13}$$

where

$$\delta := \delta(\omega) \equiv \rho^{\frac{1}{2}}(\Delta^T(\omega)\Delta(\omega)), \tag{14}$$

$$T(t) := T(\omega, t) \equiv (1-\omega)^2(1+\omega^2 t) + \omega^2 t + (1-\omega+\omega^2 t)^2, \tag{15}$$

$$C := C(\omega) \equiv (1-\omega)^4, \tag{16}$$

and where t $(= \rho^2(B) = s_1^2)$ *is the square of the spectral radius of the associated block Jacobi iteration matrix* B *in* (3) (*and also equals the square of the largest singular value of the matrix* M *in* (4)–(5)).

Proof. In case $t = 0$, we immediately have that $\delta^2 = (1-\omega)^2$ and therefore $\widehat{\delta} = \rho(B) = 0$ for $\widehat{\omega} = 1$. Thus, in what follows we shall assume that $t \in (0, 1)$.

From the structure of the matrix $\Delta(\omega)$ in (9)–(10) and the expressions in (13)–(16), we have that

$$\begin{aligned}
\delta^2 = ||\Delta(\omega)||_2^2 = \rho\left(\Delta^T(\omega)\Delta(\omega)\right) &= \max\left\{\max_{i=1,\ldots,\min\{p,q\}} ||\Delta_i(\omega)||_2^2,\ (1-\omega)^2\right\} \\
&= \max\left\{\max_{i=1,\ldots,\min\{p,q\}} \rho(\Delta_i^T(\omega)\Delta_i(\omega)),\ (1-\omega)^2\right\} \qquad (17) \\
&= \max_{i=1,\ldots,\min\{p,q\}} \frac{1}{2}\left[T(t_i) + \left[T^2(t_i) - 4C\right]^{\frac{1}{2}}\right].
\end{aligned}$$

For fixed $\omega \in [0,2]$, the function T is a strictly increasing function of $t_i = s_i^2 \in [0,1)$, $i = 1, \ldots, \min\{p,q\}$, because

$$\frac{\partial T(\omega,t)}{\partial t} = \omega^2[(2-\omega)^2 + 2\omega^2 t] \geq 0, \tag{18}$$

an observation made by Young (see pages 246–247 in [10]), which implies the monotonicity of $||\Delta(\omega,t_i)||_2^2$ as a function of $t_i = s_i^2 \geq 0$. But $||\Delta(\omega,t_i)||_2^2 \geq ||\Delta(\omega,0)||_2^2 = (1-\omega)^2$. Therefore, in view of (17)–(18) and the previous discussion, the problem of finding $\widehat{\omega} \in (0,2)$ with $||\mathcal{L}_{\widehat{\omega}}||_2 = \min_{\omega\in(0,2)} ||\mathcal{L}_\omega||_2$ boils down to minimizing in (13), namely, to the minimization of δ^2, where

$$\delta^2 = ||\Delta(\omega)||_2^2 = \rho(\Delta_1^T(\omega)\Delta_1(\omega)) = \frac{1}{2}\left[T(t) + \left[T^2(t) - 4C\right]^{\frac{1}{2}}\right] =: L(\omega) \tag{19}$$

as a function of $\omega \in (0,2)$, where $t = \rho^2(B)$ is fixed in $(0,1)$. □

We are now in a position to determine the $\widehat{\omega}$ which minimizes (13).

THEOREM 2.2. *Under the assumptions of Lemma* 2.1 *and for any fixed* $t \in (0,1)$, *the value of* ω, *call it* $\widehat{\omega}$, *which yields the minimum in* (13) *is the unique real positive root in* $(0,1)$ *of the quartic equation*

$$f(\omega) := (t^2+t^3)\omega^4 + (1-4t^2)\omega^3 + (-5+4t+4t^2)\omega^2 + (8-8t)\omega + (-4+4t) = 0. \tag{20}$$

In fact $\widehat{\omega} \in (0,\omega^*)$, *where* ω^* *is the unique real positive root in* $(0,1)$ *of the cubic*

$$g(\omega) := (t+t^2)\omega^3 - 3t\omega^2 + (1+2t)\omega - 1. \tag{21}$$

Proof. The function $L(\omega) = \delta^2(\omega)$ in (19) is differentiable in $(0,2)$. This follows immediately from the strictly increasing nature of T since

$$T^2(\omega,t) - 4C(\omega) \geq T^2(\omega,0) = 0,$$

where equality occurs only for $\omega = 0$. Thus, $L(\omega)$ attains its minimum value either at the endpoints of $(0,2)$ or at the zeros of its first derivative. For the endpoints, we have (see the second part of Lemma 2.9 in [5]) that

$$L(2) = 1 + 8t^2 + 4t\sqrt{1+4t^2} \;>\; 1 \;=\; L(0). \tag{22}$$

Also, it can be found that (see (2.20) of Lemma 2.6 in [5])

$$\lim_{\omega\to 0^+} \frac{\partial L}{\partial \omega} \;=\; -4(1-\sqrt{t}) \;<\; 0. \tag{23}$$

In view of (22) and (23) there is *no* global minimum of $L(\omega)$ at either $\omega = 0$ or $\omega = 2$.

For $\omega \in (0,2)$ elementary but lengthy computations show (see proofs of Lemmas 2.5–2.7 in [5]) that $\frac{\partial L}{\partial \omega}|_{\omega=1} = \frac{\partial T}{\partial \omega}|_{\omega=1} = 4t^2 > 0$ and so $\frac{\partial L}{\partial \omega} = 0$ is equivalent to

$$f(\omega) := t^3\omega^4 + (t^2\omega^2 + \omega - 1)(2-\omega)^2 + 4t(\omega-1)^2 = 0. \tag{24}$$

By inspection, there holds that $f(\omega) > 0$ for $\omega \in [1,2]$. Expanding (24) and rearranging terms gives (20). To prove the uniqueness of $\widehat{\omega} \in (0,1)$, the zero of $f(\omega)$, note that $f(0) = -4(1+t) < 0$ and $f(1) = t^2(1+t) > 0$. Furthermore, to show that $f'(\omega) > 0$ for $\omega \in (0,1)$, more complicated yet elementary arguments, including Descartes' rule

of signs, are required (see proofs of Lemmas 2.7 and 2.10 in [5]). It can actually be shown that $\widehat{\omega} < \omega^* < 1$, where ω^* denotes the (unique) zero of $\frac{\partial T}{\partial \omega}$ in $(0,2)$, given by (21) (see proofs of Lemmas 2.5 and 2.6 in [5]). □

Remark. It is worth mentioning that the table on page 247 of Young's book [10] gives, among other items, the values of $\widehat{\omega}$ and the corresponding $\| \mathcal{L}_{\widehat{\omega}} \|_2$ for $\sqrt{t} = 0, .1, \ldots, 1$. According to Young, these values were found numerically. Our results in Theorem 2.2 now confirm, theoretically, Young's findings.

3. Euclidean norm minimization of the MSOR operator. We now turn to Problem II in (12) and, for $k = 1$, we completely resolve this minimization problem. On pages 283–288 of Young's book [10], the following theorem is given.

THEOREM 3.1 (see Young [10, Theorem 8.4.1]). *If A is a positive definite matrix of the form* (3) *and if the spectral radius $\rho(B)$ of B of* (3) *satisfies*

$$t := \rho^2(B) \in \left[\frac{1}{3}, 1\right), \tag{25}$$

then

$$\| \mathcal{L}_{\widehat{\omega}_1, \widehat{\omega}_2} \|_2 = \frac{1+t}{3-t}, \tag{26}$$

where

$$(\widehat{\omega}_1, \widehat{\omega}_2) = \left(\frac{4}{5+t}, \frac{4}{3-t}\right) \tag{27}$$

and, unless $\omega_1 = \widehat{\omega}_1$ and $\omega_2 = \widehat{\omega}_2$,

$$\| \mathcal{L}_{\omega_1, \omega_2} \|_2 > \| \mathcal{L}_{\widehat{\omega}_1, \widehat{\omega}_2} \|_2 . \tag{28}$$

The proof of Theorem 3.1, whose statement is practically that of Problem II, for $k = 1$, defined in (12), is given in [12] (see also [10]). However, it is partly evidential and, in any case, covers *only* the case when $t \in [\frac{1}{3}, 1)$. In this section we develop quite a different approach from that of [12] and [10] which allows us to extend the analysis to the whole interval $[0, 1)$.

We begin with Lemma 3.2 which is analogous to Lemma 2.1. The proof of our main result in Theorem 3.3, whose statement immediately follows Lemma 3.2, requires a series of lemmas which are presented here without proof. Their proofs can be found in [5].

LEMMA 3.2. *Let $A \in \mathbb{R}^{n,n}$ be a symmetric positive definite matrix with "Property A" and of the block form* (3). *Then, Problem* II *for k=1 is equivalent to the determination of the quantity*

$$\widehat{\delta}^2 := \delta^2(\widehat{\omega}_1, \widehat{\omega}_2) \equiv \min_{\omega_1, \omega_2 \in (0,2)} \delta^2$$

$$= \min_{\omega_1, \omega_2 \in (0,2)} \begin{cases} \max\{(1-\omega_1)^2,\ (1-\omega_2)^2\} & \text{if } T(0) \geq T(t), \\ \frac{1}{2}\left[T(t) + \left[T^2(t) - 4C\right]^{\frac{1}{2}}\right] & \text{if } T(t) \geq T(0), \end{cases} \tag{29}$$

where

$$\delta := \delta(\omega_1, \omega_2) \equiv \rho^{\frac{1}{2}}\left(\Delta^T(\omega_1, \omega_2)\Delta(\omega_1, \omega_2)\right) \tag{30}$$

with $\Delta(\omega_1,\omega_2)$ *given in* (9)–(10),

$$T(t) := T(\omega_1,\omega_2,t) \;\equiv\; (1-\omega_1)^2 + (1-\omega_1)^2\omega_2^2 t + \omega_1^2 t + (1-\omega_2+\omega_1\omega_2 t)^2, \tag{31}$$

$$C := C(\omega_1,\omega_2) \;\equiv\; (1-\omega_1)^2(1-\omega_2)^2, \tag{32}$$

and where t *is the square of the spectral radius of the associated block Jacobi iteration matrix* B *in* (3).

Proof. Working as in the proof of Lemma 2.1 but using the matrix $\Delta(\omega_1,\omega_2)$ given in (9)–(10) and the expression (30) for δ, we have that the characteristic equation of $\Delta^T(\omega_1,\omega_2)\Delta(\omega_1,\omega_2)$ is given by

$$\left[\lambda - (1-\omega_j)^2\right]^{|p-q|} \prod_{i=1}^{q} \left[\lambda^2 - T(s_i^2)\lambda + C\right],\; s_i^2 \in [0,t], \tag{33}$$

where $j = 1$, whenever $p \geq q$, and $j = 2$ otherwise. In (33), s_i^2 are the squares of the eigenvalues of the block Jacobi iteration matrix B. Compared to [10, equations (8.4.8)–(8.4.9), p. 284], the characteristic equation (33) has an extra factor, the leftmost. Now for $t \in [0,1)$, $T(t) = T(s^2) := T(\omega_1,\omega_2,s^2) \geq 0 \;\forall\; s^2 = s_i^2 \in [0,t]$, $i = 1,\ldots,\min\{p,q\}$. On the other hand, $\frac{\partial^2 T(s^2)}{\partial (s^2)^2} = 2\omega_1^2\omega_2^2 > 0$, implying that $\frac{\partial T(s^2)}{\partial s^2}$ is strictly increasing. Consequently, $T(s^2)$ is a convex function on $[0,t]$ whose maximum is attained at one of the endpoints of the interval $[0,t]$. Following Young [10] we introduce the notations

$$\begin{aligned} \mathrm{R} &:= \{(\omega_1,\omega_2) \in \mathrm{R} \mid (\omega_1,\omega_2) \in (0,2)\times(0,2)\}, \\ \mathrm{RI} &:= \{(\omega_1,\omega_2) \in \mathrm{RI} \mid (\omega_1,\omega_2) \in \mathrm{R} \text{ and } T(0) \geq T(t)\}, \\ \mathrm{RII} &:= \{(\omega_1,\omega_2) \in \mathrm{RII} \mid (\omega_1,\omega_2) \in \mathrm{R} \text{ and } T(t) \geq T(0)\}, \\ \Gamma &:= \mathrm{RI} \cap \mathrm{RII}. \end{aligned} \tag{34}$$

Except for the leftmost factors, the roots of each factor in the products (33) are given by the expressions

$$\lambda \;=\; \frac{1}{2}\left[T(s^2) \pm \left[T^2(s^2) - 4C\right]^{\frac{1}{2}}\right] \;\forall\; s^2 = s_i^2 \in [0,t]. \tag{35}$$

For each $s^2 = s_i^2$, the largest of the two roots is the one with the plus sign in front of the radical. Moreover, because $T(\cdot)$ attains its maximum at one of the endpoints of the interval $[0,t]$, the overall maximal root λ corresponds to the larger of $T(t)$ and $T(0)$. Since $T(0) = (1-\omega_1)^2 + (1-\omega_2)^2$, it is readily seen that for $(\omega_1,\omega_2) \in \mathrm{RI}$, the largest of the eigenvalues in (33) is given by the first expression in (29). For $(\omega_1,\omega_2) \in \mathrm{RII}$, this eigenvalue is

$$\frac{1}{2}\left[T(t) + \left[T^2(t) - 4C\right]^{\frac{1}{2}}\right] \geq \frac{1}{2}\left[T(0) + \left[T^2(0) - 4C\right]^{\frac{1}{2}}\right] = \max\{(1-\omega_1)^2, (1-\omega_2)^2\}. \tag{36}$$

Therefore, the largest of the eigenvalues in (33) is given by the second expression in (29). These results are almost identical to the results in the first part of the proof of Theorem 8.4.1 in [10]. ∎

We are now ready to state the main result of this section.

THEOREM 3.3. *Under the assumptions of Lemma* 3.2 *and for any fixed* $t \in [0,1)$*, the pair* (ω_1, ω_2)*, call it* $(\widehat{\omega}_1, \widehat{\omega}_2)$*, which yields the minimum* $\widehat{\delta}$ *in* (29) *is as follows: for* $t \in [0, \frac{1}{3}]$

$$(\widehat{\omega}_1,\ \widehat{\omega}_2) \ = \ \left(\frac{1}{1+t},\ \frac{1}{1-t}\right), \tag{37}$$

when

$$\widehat{\delta} \ = \ \sqrt{\frac{t}{1+t}}, \tag{38}$$

while for $t \in [\frac{1}{3}, 1)$

$$(\widehat{\omega}_1,\ \widehat{\omega}_2) \ = \ \left(\frac{4}{5+t},\ \frac{4}{3-t}\right), \tag{39}$$

when

$$\widehat{\delta} \ = \ \frac{1+t}{3-t}. \tag{40}$$

Note. To find which of $T(0)$ and $T(t)$ is the largest in Lemma 3.2, we consider the difference

$$D \ := \ T(t) - T(0) \ = \ \left[(1+t)\omega_2^2 + 1\right]\omega_1^2 - 2(2\omega_2 - 1)\omega_2\omega_1 + \omega_2^2 \tag{41}$$

as a function of ω_1. The discriminant $d = 4\omega_2^3\left[(3-t)\omega_2 - 4\right]$ of D is negative, zero, or positive depending on whether $\omega_2 \in (0, \frac{4}{3-t})$, $\omega_2 = \frac{4}{3-t}$, or $\omega_2 \in (\frac{4}{3-t}, 2)$, in which case D has none, one, or two real zeros, respectively. Observing that for any $t \in [0,1)$, $\omega_2 = \frac{4}{3-t}$ $(\in [\frac{4}{3}, 2])$ yields, by (41), $D = 0$ and so $\omega_1 = \frac{4}{5+t} (\in [\frac{2}{3}, \frac{4}{3}])$ and $(\omega_1, \omega_2) = \left(\frac{4}{5}, \frac{4}{3}\right) \in$ R. We conclude that neither of the two regions RI and RII is empty. To obtain an idea about the shape of the boundary curve Γ, we give the following lemma whose proof can be found in [5].

LEMMA 3.4. *For any fixed* $t \in [0,1)$ *and any* $\omega_2 \in (\frac{4}{3-t}, 2)$*, the quadratic* (41) *has two distinct real roots* $\omega_1' < \omega_1''$ *such that* $0 < \omega_1' < \omega_1'' < 2$.

In Figure 1 the curve Γ, which is the boundary between RI and RII, and its position in R is drawn for the values $t = 0.6,\ 0.75$, and 0.9. Note the similarity of the curve Γ to the graph shown in Figure 4.1 on page 285 of [10].

To complete the proof of Theorem 3.3 we must examine two basic cases, starting with case $(\omega_1, \omega_2) \in$ RI. The following lemma can be proved.

LEMMA 3.5. *For any fixed* $t \in [0,1)$ *and any* $(\omega_1, \omega_2) \in$ RI*, the solution to the optimization problem* (29) *occurs at the point* $(\widehat{\omega}_1, \widehat{\omega}_2) = (\frac{4}{5-t}, \frac{4}{3-t})$ *and the corresponding minimum value of* δ *is* $\widehat{\delta} = \frac{1+t}{3-t}$.

Proof. For the proof see Lemma 3.6 in [5]. □

We now examine the case when $(\omega_1, \omega_2) \in$ RII. Its proof is harder than the previous case and requires the following sequence of lemmas which are presented without proof.

LEMMA 3.6. *Suppose that* ω_1 *and* ω_2 *are not equal to one. Then, for any fixed* $t \in (0,1)$*, the points* $(\omega_1, \omega_2) \in$ RII *at which the root function*

$$\lambda := \lambda(\omega_1, \omega_2, t) \equiv \frac{1}{2}\left[T(t) + \left[T^2(t) - 4C\right]^{\frac{1}{2}}\right] \tag{42}$$

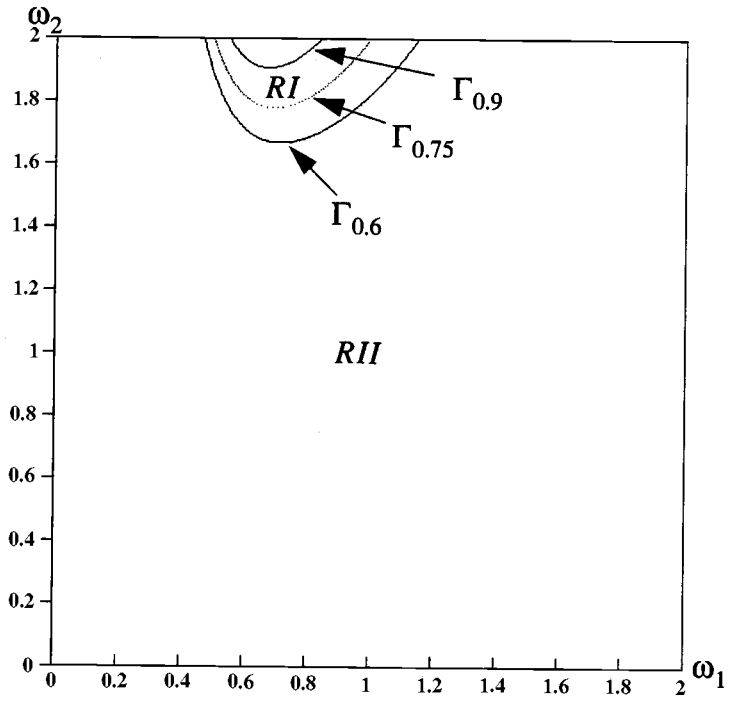


FIG. 1. *The boundary curves* Γ *for the regions* RI *and* RII.

of the quadratic

$$\lambda^2 - T(t)\lambda + C = 0 \tag{43}$$

attains its minimum value occur at stationary points of (42). *Moreover, these stationary points are the common roots of the two quadratics in* ω_1*:*

$$P_1(\omega_1) := a_1\omega_1^2 + b_1\omega_1 + c_1 = 0 \tag{44}$$

and

$$P_2(\omega_1) := a_2\omega_1^2 + b_2\omega_1 + c_2 = 0, \tag{45}$$

where

$$\begin{aligned} a_1 := a_1(\omega_2) &= (t^3 - t)\omega_2^4 + (t^2 + t)\omega_2^3 + (2t^2 + t - 1)\omega_2^2 \\ &\quad +(t + 1)\omega_2 + (t + 1), \\ b_1 := b_1(\omega_2) &= (-3t^2 + 2t + 1)\omega_2^4 + (2t^2 - 4t - 2)\omega_2^3 \\ &\quad +(-2t + 2)\omega_2^2 + (2t - 2)\omega_2, \\ c_1 := c_1(\omega_2) &= (2t - 2)\omega_2^4 + (-3t + 5)\omega_2^3 + (t - 3)\omega_2^2, \end{aligned} \tag{46}$$

and

$$\begin{aligned} a_2 := a_2(\omega_2) &= (t^2 + t)\omega_2 + (t + 1), \\ b_2 := b_2(\omega_2) &= (-t^2 + t)\omega_2^2 - 4t\omega_2 - 2, \\ c_2 := c_2(\omega_2) &= (t - 1)\omega_2^2 + 2\omega_2. \end{aligned} \tag{47}$$

Note. The value $t = 0$ is not included in the interval under consideration in the lemma because, when $t = 0$, the root function λ in (42) equals, via (31)–(32), $\max\{(1-\omega_1)^2, (1-\omega_2)^2\}$. However, when λ is so, the optimal values for $\widehat{\delta}$, $\widehat{\omega}_1$, and $\widehat{\omega}_2$ have already been found in (40) and (39), respectively.

Proof. For the proof see Lemma 3.7 in [5]. □

LEMMA 3.7. *The quadratic equations* (44) *and* (45) *of Lemma* 3.6 *share a common root if their resultant* P *vanishes, that is, if*

$$P := P(\omega_2) \equiv (a_1c_2 - a_2c_1)^2 - (a_1b_2 - a_2b_1)(b_1c_2 - b_2c_1) = 0 \tag{48}$$

or, equivalently, if

$$P := t_{11}\omega_2^{11} + t_{10}\omega_2^{10} + t_9\omega_2^9 + t_8\omega_2^8 + t_7\omega_2^7 + t_6\omega_2^6 + t_5\omega_2^5 + t_4\omega_2^4 + t_3\omega_2^3 + t_2\omega_2^2 = 0, \tag{49}$$

where

$$\begin{aligned}
t_{11} &= -t^8 + 5t^7 - 9t^6 + 5t^5 + 5t^4 - 9t^3 + 5t^2 - t,\\
t_{10} &= t^8 - 11t^7 + 34t^6 - 39t^5 + 39t^3 - 34t^2 + 11t - 1,\\
t_9 &= 2t^7 - 22t^6 + 58t^5 - 46t^4 - 26t^3 + 62t^2 - 34t + 6,\\
t_8 &= 4t^7 - 27t^6 + 40t^5 + 13t^4 - 60t^3 + 23t^2 + 16t - 9,\\
t_7 &= 20t^6 - 104t^5 + 140t^4 + 16t^3 - 148t^2 + 88t - 12,\\
t_6 &= 4t^6 + 16t^5 - 108t^4 + 112t^3 + 60t^2 - 128t + 44,\\
t_5 &= 24t^5 - 56t^4 - 16t^3 + 80t^2 - 8t - 24,\\
t_4 &= 52t^4 - 104t^3 - 16t^2 + 104t - 36,\\
t_3 &= 48t^3 - 48t^2 - 48t + 48,\\
t_2 &= 16t^2 - 16.
\end{aligned} \tag{50}$$

Furthermore, for $t \in (0,1)$, *the* 11 *roots of the resultant* (49) *are as follows:*

$$\begin{aligned}
&0, \quad 0, \quad 1, \quad -\tfrac{1}{t}, \quad \tfrac{1}{1-t}, \quad \tfrac{2}{1-t}, \quad \tfrac{2}{1-t},\\
&-\left(\tfrac{2}{1-t}\right)^{\frac{1}{2}}, \quad -\left(\tfrac{2}{1-t}\right)^{\frac{1}{2}}, \quad \left(\tfrac{2}{1-t}\right)^{\frac{1}{2}}, \quad \left(\tfrac{2}{1-t}\right)^{\frac{1}{2}}.
\end{aligned} \tag{51}$$

Moreover, the distinct values of $\omega_2 \in (0,2) \setminus \{1\}$ *given in* (51) *which are admissible as ordinates of possible stationary points of the function* λ *of Lemma* 3.6 *are the following:*

$$\omega_2 = \frac{1}{1-t}, \ \left(\frac{2}{1-t}\right)^{\frac{1}{2}} \quad \forall\ t \in \left(0, \frac{1}{2}\right). \tag{52}$$

Proof. For the proof see Lemmas 3.8–3.10 in [5]. □

LEMMA 3.8. *The common roots of the quadratics* (44) *and* (45) *in Lemma* 3.6 *given by the expression*

$$-\frac{a_1c_2 - a_2c_1}{a_1b_2 - a_2b_1} \tag{53}$$

are the following:

$$\begin{cases}
(\omega_1, \omega_2)_1 = \left(\frac{1}{1+t}, \frac{1}{1-t}\right) & \forall\ t \in \left(0, \frac{1}{2}\right),\\
(\omega_1, \omega_2)_2 = \left(\frac{2(5-3t)-(7-t)(2-2t)^{\frac{1}{2}}}{(1+t)\left[3-t-2(2-2t)^{\frac{1}{2}}\right]}, \left(\frac{2}{1-t}\right)^{\frac{1}{2}}\right) & \forall\ t \in \left(0, \frac{1}{2}\right).
\end{cases} \tag{54}$$

Moreover, of the above points only

$$(\omega_1, \omega_2)_1 \;=\; \left(\tfrac{1}{1+t}, \tfrac{1}{1-t}\right) \;\forall\; t \in (0, \tfrac{1}{3}] \tag{55}$$

*lies in R*II.

Proof. For the proof see Lemma 3.11 in [5]. ∎

We are now ready for the proof of Theorem 3.3.

Proof of Theorem 3.3. We begin with the case where ω_1 and/or ω_2 are different than one, a case that was excluded from the last few lemmas. From the results of Lemmas 3.6–3.8, it is clear that we need to determine the value of λ in (42) for the pair $(\omega_1, \omega_2)_1$ in (55). For this we are required to evaluate $T(\omega_1, \omega_2, t)$ and $C(\omega_1, \omega_2)$ at $(\omega_1, \omega_2)_1$ for all $t \in (0, \frac{1}{3}]$. It can be found that

(56)
$$T\left(\frac{1}{1+t}, \frac{1}{1-t}, t\right) = \frac{t(1-2t+2t^2)}{(1+t)(1-t)^2} \text{ and } C\left(\frac{1}{1+t}, \frac{1}{1-t}\right) \;=\; \frac{t^4}{(1+t)^2(1-t)^2},$$

implying that

$$\lambda \;=\; \frac{t}{1+t} \;\; \forall\; t \in \left(0, \frac{1}{3}\right]. \tag{57}$$

To complete the proof we need to consider the cases where ω_1 and/or ω_2 are equal to one. Starting with $\omega_1 = 1$, we find that $D \;=\; T(t) - T(0) \;>\; 0$ if and only if $\omega_2 \in (0, \frac{1+\sqrt{3-t}}{2-t})$ and $t \in (0, \frac{3}{4})$. Minimizing $T(t)$ as function of ω_2 yields that $T(t)$ attains a minimum at $\omega_2 = \frac{1}{1-t}$ provided that $t \in (0, \frac{1}{2})$. For $t \in [\frac{1}{2}, \frac{3}{4})$, $\min T(t) \geq t$, for all $\omega_2 \in (0, 2)$. Therefore,

$$T\left(1, \frac{1}{1-t}, t\right) \;=\; t \;>\; \lambda \;=\; \frac{t}{1+t} \quad \forall\; t \in \left(0, \frac{1}{3}\right].$$

On the other hand and in view of Lemma 3.5, it can be found that

$$T\left(1, \omega_2, t\right) \;\geq\; t \;>\; \left(\frac{1+t}{3-t}\right)^2 \quad \forall\; \omega_2 \in (0, 2) \text{ and } \forall\; t \in \left(\frac{1}{3}, \frac{3}{4}\right).$$

Suppose next that $\omega_2 = 1$. Then we find that $D \;=\; T(t) - T(0) \;>\; 0$. Minimizing $T(t)$ as function of ω_1 yields that $T(t)$ attains a minimum at $\omega_1 = \frac{1}{1+t}$ for all $t \in (0, 1)$. Therefore,

$$T\left(\frac{1}{1+t}, 1, t\right) \;=\; t \;>\lambda \;=\; \frac{t}{1+t} \;\; \forall\; t \in \left(0, \frac{1}{3}\right],$$

while from Lemma 3.5 we have that

$$T\left(\frac{1}{1+t}, 1, t\right) \;=\; t \;>\; \left(\frac{1+t}{3-t}\right)^2 \quad \forall\; t \in \left[\frac{1}{3}, 1\right).$$

This, together with the results of Lemma 3.5, completes the theorem's proof. ∎

4. Energy norm minimization of the SOR and MSOR operators. Under the assumptions of Lemmas 2.1 and 3.2 we give two theorems below concerning the energy norms of the SOR and MSOR operators which can actually be found in Young's book [10] but seem not to be well known among the researchers working in the area. We include them here for completeness and for the sake of comparison with the results obtained in sections 2 and 3. Also, based on the reductions in (9)–(10) obtained by Golub and dePillis in [1], we were able to give a much simpler proof of Theorem 15.2.1 in [10] which can be found in Theorem 4.1 in [5].

THEOREM 4.1. *Under the assumptions of Lemma* 2.1*,*

$$\min_{\omega\in(0,2)} \| \mathcal{L}_\omega\|_{A^{\frac{1}{2}}} = \min_{\omega\in(0,2)} \| A^{\frac{1}{2}}\mathcal{L}_\omega A^{-\frac{1}{2}}\|_2 = \| \mathcal{L}_1\|_{A^{\frac{1}{2}}} = \rho(B), \tag{58}$$

where B is the block Jacobi iteration matrix associated with A.

COROLLARY 4.2. *Under the assumptions of Lemma* 2.1*,*

$$\min_{\omega\in(0,2)} \| \mathcal{L}_\omega \|_{A^{\frac{1}{2}}} = \| \mathcal{L}_1 \|_{A^{\frac{1}{2}}} = \rho(B) \le \| \mathcal{L}_{\widehat{\omega}} \|_2 = \min_{\omega\in(0,2)} \| \mathcal{L}_\omega \|_2, \tag{59}$$

where $\widehat{\omega}$ is the value of the optimal ω of Theorem 2.2 *and where equality in* (59) *holds only in the trivial case when $\sqrt{t} = \rho(B) = 0$.*

As is seen, the energy norm of the SOR operator gives a better minimum value than that of the Euclidean norm.

For the energy norm of the corresponding MSOR operator we simply state part of Theorem 8.5.1 in [10]. It is based on results in [11] and [9].

THEOREM 4.3. *Under the assumptions of Lemma* 3.2*,*

$$\min_{\omega_1,\omega_2\in(0,2)} \| \mathcal{L}_{\omega_1,\omega_2} \|_{A^{\frac{1}{2}}} = \| \mathcal{L}_{1,1} \|_{A^{\frac{1}{2}}} = \rho(B). \tag{60}$$

COROLLARY 4.4. *Under the assumptions of Lemma* 3.2*,*

$$\begin{aligned} &\min_{\omega_1,\omega_2\in(0,2)} \| \mathcal{L}_{\omega_1,\omega_2} \|_{A^{\frac{1}{2}}} = \| \mathcal{L}_{1,1} \|_{A^{\frac{1}{2}}} = \rho(B) \\ &\ge \min_{\omega_1,\omega_2\in(0,2)} \| \mathcal{L}_{\omega_1,\omega_2} \|_2 = \| \mathcal{L}_{\widehat{\omega}_1,\widehat{\omega}_2} \|_2, \end{aligned} \tag{61}$$

where $(\widehat{\omega}_1,\widehat{\omega}_2)$ is the pair (37) *or* (39)*, whichever applies, and where equality in* (61) *holds only in the trivial case when $\sqrt{t} = \rho(B) = 0$.*

As is seen, the values of the minimum energy norms of the SOR and MSOR operators are identically the same and this value is larger than the minimum value of the Euclidean norm of the MSOR operator. Consequently, of the four minimum values presented in this work, and more specifically in Theorems 2.2, 3.3, 4.1, and 4.3, the best minimum is that which was found in Theorem 3.3. We shall have more to say about this in the next section.

5. Concluding remarks. We believe that, in part, the question which was raised by Golub and dePillis and which was reiterated at the beginning of the paper was motivated by a phenomenon in SOR theory called the "hump." This phenomenon occurs when an eigenvalue ν of the SOR iteration matrix whose modulus is equal to the spectral radius has a nonlinear (quadratic) elementary divisor and it can also result from the relatively large distance of the SOR iteration matrix from a normal matrix. This can cause the relative error for a small number of iterations m to actually increase since then the convergence is governed by the term $m|\nu|^{m-1}$. It is in

such a situation when it might, in fact, become beneficial to begin the iteration using relaxation parameters which are not optimal for the spectral radius. For a discussion of the hump phenomenon see Chapter 7.1 in Young's book [10].

In section 4 of [1] there is a numerical example regarding the MSOR iteration operator associated with the matrix

$$A \;=\; \text{tridiag}\left(-\frac{1}{2}, 1, -\frac{1}{2}\right) \in \mathbb{R}^{100,100}.$$

It is well known that here $t^{1/2} = \rho(B) = \cos(\pi/101) \approx .999516282$. Using numerical minimization Golub and dePillis obtain that the pair (ω_1, ω_2) which minimizes $||\mathcal{L}^{50}_{\omega_1,\omega_2}||_2$ is given by

$$(\widetilde{\omega}_1, \widetilde{\omega}_2) \approx (0.6961, 2.0000) \tag{62}$$

and the corresponding value of the Euclidean norm of $\mathcal{L}^{50}_{\widetilde{\omega}_1,\widetilde{\omega}_2}$ is given by

$$||\mathcal{L}^{50}_{\widetilde{\omega}_1,\widetilde{\omega}_2}||_2 \approx 0.9508. \tag{63}$$

For this particular example, using the optimal results of the present work in Theorem 3.3, which in this case coincide with the results in [10] because $\rho^2(B) > \frac{1}{3}$, we have

$$(\widehat{\omega}_1, \widehat{\omega}_2) \approx (0.666774151, 1.999033267). \tag{64}$$

But then

$$||\mathcal{L}_{\widehat{\omega}_1,\widehat{\omega}_2}||_2 \approx 0.999033267 \tag{65}$$

which gives that

$$||\mathcal{L}_{\widehat{\omega}_1,\widehat{\omega}_2}||_2^{50} \;\approx\; 0.9528. \tag{66}$$

We see that after 50 iterations the optimal result found numerically in [1] for the Euclidean norm of the 50th power of the MSOR iteration operator is only 0.21% better than the 50th power of the optimal Euclidean norm of the first(!) power of the MSOR iteration operator found theoretically. This might suggest that at least in some cases of practical interest, where the values of $\rho(B)$ are close to one, it would be better to minimize the Euclidean norm of the MSOR iteration operator based on a numerical approximation obtained for $\rho(B)$ rather than to estimate the optimal Euclidean norm of the kth power of the same operator for *large* k. This is because the latter minimization has to be done computationally and so the extra number of calculations may well outweigh the gain by only a slight improvement in the reduction factor.

Consider the example on page 88 of Young's book [10]. There $\rho(\mathcal{L}_{\widehat{\omega}}) = 0.8$, from which we can find out that $t = \rho^2(B) = \frac{80}{81} \approx 0.987654321$. From either Theorem 3.1 (which is Young's) or our Theorem 3.3 we find that

$$(\widehat{\omega}_1, \widehat{\omega}_2) \;\approx\; (0.668041237, 1.987730061)$$

and so

$$||\mathcal{L}_{\widehat{\omega}_1,\widehat{\omega}_2}||_2 \;=\; \frac{1+t}{3-t} \;\approx\; 0.987730061.$$

On numerically solving the inequality

$$\left(||\mathcal{L}_{\widehat{\omega}_1,\widehat{\omega}_2}||_2\right)^m - m\ \left(\rho\left(\mathcal{L}_{\omega_{1,opt},\omega_{2,opt}}\right)\right)^{m-1}\ < 0,$$

where $\omega_{1,opt}$ and $\omega_{2,opt}$ are the relaxation parameters which give the spectral radius of the MSOR operator a minimum (and which, in this case, according to Young are equal to the common value $\frac{2}{1+\sqrt{1-t}}$ and yield $\rho\left(\mathcal{L}_{\omega_{1,opt},\omega_{2,opt}}\right) = 0.8$), we find that the inequality holds for $m \leq 13$. This says then that, in the case of a hump, we should start with 13 iterations or so using the MSOR iteration operator with the relaxation parameters given in (39). Further experiments that we have carried out on examples in which the spectral radius of the Jacobi matrix is *even closer* to one show that even more iterations should be initially performed using the MSOR iteration operator when its Euclidean norm is minimal *before* switching to the SOR or MSOR iteration operators whose spectral radius is optimal. Thus, in situations when the value of $\rho(B)$ is not available precisely but is known to be very close to one, (39) tells us that the optimal pair $(\widehat{\omega}_1, \widehat{\omega}_2)$ is very close to $\left(\frac{2}{3}, 2\right)$. We therefore suggest performing initially 15 to 20 iterations using, for example, $(\widehat{\omega}_1, \widehat{\omega}_2) = (0.667, 1.99)$, before switching to an adaptive SOR method (see, e.g., Hageman and Young [3]).

Concerning the viability of starting the iterations with the SOR operator with the relaxation parameter giving its Euclidean norm a minimum as found in Theorem 2.2 or doing the same with the MSOR operator, when a Jacobi iteration matrix has a spectral radius $\rho^2(B) < \frac{1}{3}$, with the relaxation parameters chosen to give its Euclidean norm a minimum as found in Theorem 3.3, our numerical experiments indicate a poor advantage in speeding up the convergence using the above approach.

Acknowledgment. The authors express their thanks to Professor Michael Eiermann and the two referees for their constructive criticism and valuable suggestions toward the improvement of the presentation of the results of this paper.

REFERENCES

[1] G. H. Golub and J. dePillis, *Toward an effective two-parameter method*, in Iterative Methods for Large Linear Systems, D. R. Kincaid and L. Hayes, eds., Academic Press, New York, 1990, pp. 107–118.

[2] G. H. Golub and C. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[3] L. A. Hageman and D. M. Young, *Applied Iterative Methods*, Academic Press, New York, NY, 1981

[4] C. Lanczos, *Linear Differential Operators*, Van Nostrand, New York, NY, 1961.

[5] A. Hadjidimos and M. Neumann, *On the Minimization of the ℓ_2–norms of the SOR and the MSOR Operators*, CSD TR-96-012, Department of Computer Sciences, Purdue University, West Lafayette, IN, 1996.

[6] R. S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.

[7] D. M. Young, *Iterative methods for solving partial differential equations of elliptic type*, Trans. Amer. Math. Soc., 76 (1954), pp. 92–111.

[8] D. M. Young, *Convergence properties of the symmetric and unsymmetric successive overrelaxation methods and related methods*, Math. Comp., 24 (1970), pp. 793–807.

[9] D. M. Young, *Generalizations of "Property A" and Consistent Orderings*, Report CNA-6, Center for Numerical Analysis, University of Texas, Austin, TX, 1970.

[10] D. M. Young, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1971.

[11] D. M. Young and D. R. Kincaid, *Norms of the Successive Overrelaxation and Related Methods*, Report TNN-94, Computation Center, University of Texas, Austin, TX, 1969.

[12] D. M. Young, M. F. Wheeler, and J. Downing, *On the use of the modified successive overrelaxation method with several relaxation factors*, in Proceedings of International Federation for Information Processing 65, 1965, pp. 177–182.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 205–225, January 1998

014

USING THE MATRIX SIGN FUNCTION TO COMPUTE INVARIANT SUBSPACES*

ZHAOJUN BAI† AND JAMES DEMMEL‡

Abstract. The matrix sign function has several applications in system theory and matrix computations. However, the numerical behavior of the matrix sign function, and its associated divide-and-conquer algorithm for computing invariant subspaces, are still not completely understood. In this paper, we present a new perturbation theory for the matrix sign function, the conditioning of its computation, the numerical stability of the divide-and-conquer algorithm, and iterative refinement schemes. Numerical examples are also presented. An extension of the matrix-sign-function-based algorithm to compute left and right deflating subspaces for a regular pair of matrices is also described.

Key words. matrix sign function, Newton's method, eigenvalue problem, invariant subspace, deflating subspaces

AMS subject classifications. 65F15, 65F35, 65F30, 15A18

PII. S0895479896297719

1. Introduction.

Since the matrix sign function was introduced in the early 1970s, it has been the subject of numerous studies and used in many applications. For example, see [30, 31, 11, 26, 23] and references therein. Our main interest here is to use the matrix sign function to build parallel algorithms for computing invariant subspaces of nonsymmetric matrices, as well as their associated eigenvalues. It is a challenge to design a parallel algorithm for the nonsymmetric eigenproblem that uses coarse grain parallelism effectively, scales for larger problems on larger machines, does not waste time dealing with the parts of the spectrum in which the user is not interested, and deals with highly nonnormal matrices and strongly clustered spectra. In the work of [2], after reviewing the existing approaches, we proposed a design of a parallel nonsymmetric eigenroutine toolbox, which includes the basic building blocks (such as LU factorization, matrix inversion, and the matrix sign function), standard eigensolver routines (such as the QR algorithm), and new algorithms (such as spectral divide-and-conquer using the sign function). We discussed how these tools could be used in different combinations on different problems and architectures, for extracting all or some of the eigenvalues of a nonsymmetric matrix, and/or their corresponding invariant subspaces. Rather than using "black box" eigenroutines such as provided by EISPACK [32, 21] and LAPACK [1], we expect the toolbox approach to allow us more flexibility in developing efficient problem-oriented eigenproblem solvers on high-performance machines, especially on parallel distributed memory machines.

* Received by the editors January 25, 1996; accepted for publication (in revised form) by V. Mehrmann February 4, 1997.

http://www.siam.org/journals/simax/19-1/29771.html

† Department of Mathematics, University of Kentucky, Lexington, KY 40506 (bai@ms.uky.edu). The research of this author was supported in part by ARPA grant DM28E04120 and P-95006 via a subcontract from Argonne National Laboratory, by NSF grant ASC-9313958, and in part by DOE grant DE-FG03-94ER25219 via subcontracts from the University of California at Berkeley.

‡ Computer Science Division and Mathematics Department, University of California, Berkeley, CA 94720 (demmel@cs.berkeley.edu). The research of this author was funded in part by ARPA contract DAAH04-95-1-0077 through University of Tennessee subcontract ORA7453.02, ARPA contract DAAL03-91-C-0047 through University of Tennessee subcontract ORA4466.02, NSF contracts ASC-9313958 and ASC-9404748, DOE contracts DE-FG03-94ER25219, DE-FG03-94ER25206, DOE contract W-31-109-Eng-38 through subcontract 951322401 with Argonne National Laboratory, and NSF Infrastructure Grants CDA-8722788 and CDA-9401156.

However, the numerical accuracy and stability of the matrix sign function and divide-and-conquer algorithms based on it are poorly understood. In this paper, we will address these issues. Much of this work also appears in [3].

Let us first restate some of basic definitions and ideas to establish notation. The matrix sign function of a matrix A is defined as follows [30]: let

$$A = X\mathrm{diag}(J_+, J_-)X^{-1}$$

be the Jordan canonical form of a matrix $A \in \mathbf{C}^{n\times n}$, where the eigenvalues of J_+ lie in the open right half-plane ($\mathbf{C}_+$) and those of J_- lie in the open left half-plane ($\mathbf{C}_-$). Then the matrix sign function of A is

$$\mathrm{sign}(A) = X\mathrm{diag}(I, -I)X^{-1}.$$

We assume that no eigenvalue of A lies on the imaginary axis; otherwise, $\mathrm{sign}(A)$ is not defined. It is easy to show that the spectral projection corresponding to the eigenvalues of A in the open right and left half-planes are $P_\pm = \frac{1}{2}(I \pm \mathrm{sign}(A))$, respectively. Let the leading columns of an orthogonal matrix Q span the range space of P_+ (for example, Q may be computed by the rank-revealing QR decomposition of P_+). Then Q yields the spectral decomposition

$$Q^T A Q = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \tag{1}$$

where $\lambda(A_{11})$ are the eigenvalues of A in $\mathbf{C}_+$, and $\lambda(A_{22})$ are the eigenvalues of A in $\mathbf{C}_-$. The algorithm proceeds in a divide-and-conquer fashion by computing the eigenvalues of A_{11} and A_{22}.

Rather than using the Jordan canonical form to compute $\mathrm{sign}(A)$, it can be shown that $\mathrm{sign}(A)$ is the limit of the following Newton iteration:

$$A_{k+1} = \frac{1}{2}(A_k + A_k^{-1}) \quad \text{for} \quad k = 0, 1, 2, \ldots, \quad \text{with} \quad A_0 = A. \tag{2}$$

The iteration is globally and ultimately quadratic convergent. There exist different scaling schemes to speed up the convergence of the iteration, and make it more suitable for parallel computation. By computing the matrix sign function of a Möbius transformation of A, the spectrum can be divided along arbitrary lines and circles, rather than just along the imaginary axis. See report [2] and references therein for more details.

Unfortunately, in finite precision arithmetic, the ill conditioning of a matrix A_k with respect to inversion and rounding errors may destroy the convergence of the Newton iteration (2) or cause convergence to the wrong answer. Consequently, the left bottom corner block of the matrix $Q^T A Q$ in (1) may be much larger than $\mathbf{u}\|A\|$, where $\mathbf{u}$ denotes machine precision. This means that it is not numerically stable to approximate the eigenvalues of A by the eigenvalues of A_{11} and A_{22}, as we would like.

In this paper, we will first study the perturbation theory of the matrix sign function, its conditioning, and the numerical stability of the overall divide-and-conquer algorithm based on the matrix sign function. We realize that it is very difficult to give a complete and clear analysis. We only have a partial understanding of when we can expect the Newton iteration to converge and how accurate it is. In a coarse analysis, we can also bound the condition numbers of intermediate matrices in the Newton iteration. Artificial and possibly very pathological test matrices are constructed to verify

our theoretical analysis. Besides these artificial tests, we also test a large number of eigenvalue problems of random matrices, and a few eigenvalue problems from applications, such as electrical power system analysis, numerical simulation of chemical reactions, and aerodynamics stability analysis. Through these examples, we conclude that the most bounds for numerical sensitivity and stability of matrix sign function computation and its based algorithms are reachable for some very pathological cases, but they are often very pessimistic. The worst cases happen rarely.

In addition, we discuss iterative refinement of an approximate invariant subspace and outline an extension of the matrix-sign-function-based algorithms to compute both left and right deflating subspaces for a regular matrix pencil $A - \lambda B$.

The rest of this paper is organized as follows. Section 2 presents a new perturbation bound for the matrix sign function. Section 3 discusses the numerical conditioning of the matrix sign function. The backward error analysis of computed invariant subspace and remarks on the matrix-sign-function-based algorithm versus the QR algorithm are presented in section 4. Section 5 presents some numerical examples for the analysis of sections 2, 3, and 4. Section 6 describes the iteration refinement scheme to improve an approximate invariant subspace. Section 7 outlines an extension of the matrix-sign-function-based algorithms for the generalized eigenvalue problem. Concluding remarks are presented in section 8.

2. A perturbation bound for the matrix sign function. When a matrix A has eigenvalues on the pure-imaginary axis, its matrix sign function is not defined. In other words, the set of *ill-posed problems* for the matrix sign function is the set of matrices with at least one pure-imaginary eigenvalue. Computationally, we have observed that when there are the eigenvalues of A close to the pure-imaginary axis, the Newton iteration and its variations are very slowly convergent and may be misconvergent. Moreover, even when the iteration converges, the error in the computed matrix sign function could be too large to use. It is desirable to have a perturbation analysis of the matrix sign function related to the distance from A to the nearest ill-posed problem.

Perturbation theory and condition number estimation of the matrix sign function are discussed in [25, 23, 29]. However, none of the existing error bounds explicitly reveals the relationship between the sensitivity of the matrix sign function and the distance to the nearest ill-posed problem. In this section, we will derive a new perturbation bound which explicitly reveals such relationship. We will denote all the eigenvalues of A with positive real part by $\lambda_+(A)$, i.e., $\lambda_+(A) = \{\lambda | \lambda \in \lambda(A), \Re(\lambda) > 0\}$. $\sigma_{\min}(A)$ denotes the smallest singular value of A. In addition, we recall the well-known inequality

$$\|(I - X)^{-1}\| \leq \frac{1}{1 - \|X\|} \quad \text{if } \|X\| < 1, \tag{3}$$

where $\|\cdot\|$ is the matrix 2-norm.

THEOREM 2.1. *Suppose A has no pure-imaginary and zero eigenvalues, $A + \delta A$ is a perturbation of A, and $\epsilon \equiv \|\delta A\|$. Let*

$$\omega = \max_{\tau \in \mathbb{R}} \|(i\tau I - A)^{-1}\| = \frac{1}{\min\limits_{\tau \in \mathbb{R}} \sigma_{\min}(i\tau I - A)} \equiv \frac{1}{d_A}. \tag{4}$$

Then

$$\|\text{sign}(A)\| \leq \frac{4}{\pi}\omega\|A\| + 3. \tag{5}$$

Furthermore, if

$$\omega\epsilon < 1, \tag{6}$$

then

$$\|\mathrm{sign}(A+\delta A) - \mathrm{sign}(A)\| \le \frac{4}{\pi}\frac{\omega^2\epsilon}{1-\omega\epsilon}(\|A\|+\epsilon) + 2\frac{\epsilon}{\|A\|}. \tag{7}$$

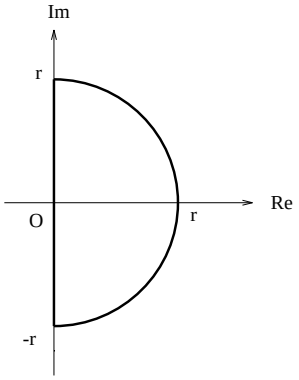


FIG. 1. *The semicircle* Γ.

Proof. We only prove the bound (7). The bound (5) can be proved by using a similar technique. Following Roberts [30] (or Kato [24]), the matrix sign function can also be defined using Cauchy integral representation:

$$\mathrm{sign}(A) = 2\,\mathrm{sign}^+(A) - I, \tag{8}$$

where

$$\mathrm{sign}^+(A) = \frac{1}{2\pi i}\int_\Gamma (\zeta I - A)^{-1} d\zeta,$$

Γ is any simple closed curve with positive direction enclosing $\lambda_+(A)$. $\mathrm{sign}^+(A)$ is the spectral projector for $\lambda_+(A)$. Here, without loss of generality, we take Γ to be a semicircle with radius $r = 2\max\{\|A\|, \|A+\delta A\|\}$ (see Figure 1). From the definition (8) of $\mathrm{sign}(A)$, it is seen that to study the stability of the matrix sign function of A to the perturbation δA, it is sufficient to just study the sensitivity of the projection $\mathrm{sign}^+(A)$.

Let $\mathrm{sign}^+(A+\delta A)$ be the projection corresponding to $\lambda_+(A+\delta A)$, from the condition (6), no eigenvalues of A are perturbed across or on the pure imaginary axis, and the semicircle Γ also encloses $\lambda_+(A+\delta A)$. Therefore, we have

$$\begin{aligned}
\mathrm{sign}^+(A+\delta A) - \mathrm{sign}^+(A) &= \frac{1}{2\pi i}\int_\Gamma [(\zeta I - A - \delta A)^{-1} - (\zeta I - A)^{-1}] d\zeta \\
&= \frac{1}{2\pi i}\int_{-r}^{r} [(i\tau I - A - \delta A)^{-1} - (i\tau I - A)^{-1}] i d\tau \\
&\quad + \frac{1}{2\pi i}\int_{-\pi/2}^{\pi/2} [(re^{i\theta} I - A - \delta A)^{-1} - (re^{i\theta} I - A)^{-1}] ire^{i\theta} d\theta \\
&\equiv \mathcal{I}_1 + \mathcal{I}_2,
\end{aligned}$$

where the first integral, denoted $\mathcal{I}_1$, is the integral over the straight line of the semicircle Γ, the second integral, denoted $\mathcal{I}_2$, is the integral over the curved part of the semicircle Γ. Now, by taking the spectral norm of the first integral term, and noting the definition of ω, the condition (6), and the inequality (3), we have

$$\begin{aligned}
\|\mathcal{I}_1\| &\le \frac{1}{2\pi}\int_{-r}^{r} \|[(i\tau I - A - \delta A)^{-1} - (i\tau I - A)^{-1}]\|\ |d\tau| \\
&= \frac{1}{2\pi}\int_{-r}^{r} \|[(i\tau I - A - \delta A)^{-1}\delta A(i\tau I - A)^{-1}]\|\ |d\tau| \\
&= \frac{1}{2\pi}\int_{-r}^{r} \|(I - (i\tau I - A)^{-1}\delta A)^{-1}(i\tau I - A)^{-1}\delta A(i\tau I - A)^{-1}\|\ |d\tau| \\
&\le \frac{1}{2\pi}\int_{-r}^{r} \frac{\|(i\tau I - A)^{-1}\|^2\|\delta A\|}{1 - \|(i\tau I - A)^{-1}\delta A\|}|d\tau| \\
&\le \frac{1}{2\pi}\frac{\omega^2\|\delta A\|}{1 - \omega\|\delta A\|}\ 2r.
\end{aligned}$$

By taking the spectral norm of the second integral term $\mathcal{I}_2$, we have

$$\begin{aligned}
\|\mathcal{I}_2\| &\le \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \|(re^{i\theta}I - A - \delta A)^{-1}\delta A(re^{i\theta}I - A)^{-1}\|\ r\ |d\theta| \\
&\le \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} \left\|\left(I - \frac{A+\delta A}{re^{i\theta}}\right)^{-1}\right\|\ \|\delta A\|\ \left\|\left(I - \frac{A}{re^{i\theta}}\right)^{-1}\right\| \frac{1}{r}|d\theta| \\
&\le \frac{1}{2\pi}\left(\frac{1}{1 - \|A+\delta A\|/r}\right)\ \|\delta A\|\ \left(\frac{1}{1 - \|A\|/r}\right)\ \frac{1}{r}\ \pi \\
&\le \frac{2\|\delta A\|}{r} \le \frac{\|\delta A\|}{\|A\|},
\end{aligned}$$

where the third inequality follows from (3) and the fourth follows from the choice of the radius r of the semicircle Γ. The desired bound (7) follows from the bounds on $\|\mathcal{I}_1\|$ and $\|\mathcal{I}_2\|$ and the identity

$$\operatorname{sign}(A + \delta A) - \operatorname{sign}(A) = 2(\operatorname{sign}^+(A + \delta A) - \operatorname{sign}^+(A)). \qquad \square$$

A few remarks are in order:

1. In the language of pseudospectra [35], the condition (6) means that the $\|\delta A\|$-pseudospectra of A do not cross the pure-imaginary axis.
2. From the perturbation bound (7), we see that the stability of the matrix sign function to the perturbation requires not only the $\|\delta A\|$-pseudospectra of the A to be bounded away from the pure-imaginary axis but also $\omega^2 = 1/d_A^2$ to be small (recall that d_A is the distance from A to the nearest matrix with a pure-imaginary eigenvalue).
3. It is natural to take $\omega^2 = 1/d_A^2$ as the condition number of the matrix sign function. Algorithms for computing d_A and related problems can be found in [14, 9, 8, 12].
4. The bound (7) is similar to the bound of the norm of the Fréchet derivative of the matrix sign function of A at X given by Roberts [30]:

$$\|\mathcal{F}(\operatorname{sign}(A), X)\| \le \frac{l_\Gamma}{2\pi}\left(\max_{\zeta \text{ on } \Gamma} \|(\zeta I - A)^{-1}\|^2\right)\|X\|,$$

where l_Γ is the length of the closed contour Γ.

Recently, an asymptotic perturbation bound of sign(A) was given by Byers, He, and Mehrmann [13]. They show that to first order in δA

$$\|\text{sign}(A+\delta A)-\text{sign}(A)\| \le \frac{4}{\delta}\left(1+\frac{\|A_{12}\|}{\delta}\right)^2\|\delta A\|, \tag{9}$$

where A is assumed to have the form of (1), $\|\delta A\|$ is sufficiently small, and

$$\delta = \text{sep}(A_{11},A_{22}) = \sigma_{\min}(I\otimes A_{11}-A_{22}^T\otimes I), \tag{10}$$

the separation of the matrices A_{11} and A_{22} [33]. $\otimes$ is the Kronecker product. Comparing the bounds (7) and (9), we note that first the bound (7) is a global bound and (9) is an asymptotic bound. Second, the assumption (6) for the bound (7) has a simple geometric interpretation (see remark 2 above). It is unspecified how to interpret the assumption on sufficiently small $\|\delta A\|$ for the bound (9).

3. Conditioning of matrix sign function computation. In [2], we point out that it may be much more efficient to compute $S = \text{sign}(A)$ to half-machine precision only, i.e., to compute S with an absolute error bounded by $\mathbf{u}^{1/2}\|S\|$. To avoid ill conditioning in the Newton iteration and achieve the half-machine precision, we believe that the matrix A must have condition number less than $\mathbf{u}^{-1/2}$. If A is ill conditioned, say having singular values less than $\mathbf{u}^{1/2}\|A\|$, we need to use a preprocessing step to deflate small singular values by a unitary similarity transformation, and obtain a submatrix having condition number less than $\mathbf{u}^{-1/2}$, and then compute the matrix sign function of this submatrix. Such a deflation procedure may be also needed for the intermediate matrices in the Newton iteration in the worst case.

We now look more closely at the situation of near convergence of the Newton iteration and relate the error to the distance to the nearest ill-posed problem [18]. As before, the ill-posed problems are those matrices with pure-imaginary eigenvalues. Without loss of generality, let us assume A is of the form

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \tag{11}$$

where $\lambda(A_{11}) \in \mathbf{C}_+$ and $\lambda(A_{22}) \in \mathbf{C}_-$. Otherwise, for any matrix B, by the Schur decomposition, we can write $B = Q^H A Q$, where A has the above form, and then $\text{sign}(B) = Q^H\text{sign}(A)Q$. Let R be the solution of the Sylvester equation

$$A_{11}R - RA_{22} = -A_{12}, \tag{12}$$

which must exist and be unique since A_{11} and A_{22} have no common eigenvalues. Then it is known that the spectral projector P corresponding to the eigenvalues of A_{11} is

$$P = \begin{pmatrix} I & R \\ 0 & 0 \end{pmatrix},$$

and $\|P\| = \sqrt{1+\|R\|^2}$. The following lemma relates R and the norm of the projection P to sign(A) and its condition number.

LEMMA 3.1. *Let A and R be as above. Let $\rho = \|R\| + \sqrt{1+\|R\|^2}$. Then*

1. $S \equiv \text{sign}(A) = \begin{pmatrix} I & -2R \\ 0 & -I \end{pmatrix}$.

2. $\|S\| = \|S^{-1}\| = \rho$, *and, therefore,* $\kappa(S) = \rho^2$.

Proof.

1. Let $X = \begin{pmatrix} I & R \\ 0 & I \end{pmatrix}$. It is easy to verify that if R satisfies (12), then $X^{-1}AX = \operatorname{diag}(A_{11}, A_{22})$. Therefore,
$$\begin{aligned}\operatorname{sign}(A) &= \operatorname{sign}(X\operatorname{diag}(A_{11}, A_{22})X^{-1}) = X\operatorname{sign}(\operatorname{diag}(A_{11}, A_{22}))X^{-1} \\ &= X\operatorname{diag}(I, -I)X^{-1} = \begin{pmatrix} I & -2R \\ 0 & -I \end{pmatrix}.\end{aligned}$$
2. Using the singular value decomposition (SVD) of R $URV^H = \Sigma = \operatorname{diag}(\sigma_i)$, one can reduce computing the SVD of S to computing the SVD of
$$\begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} S \begin{pmatrix} U^H & 0 \\ 0 & V^H \end{pmatrix} = \begin{pmatrix} I & -2\Sigma \\ 0 & -I \end{pmatrix}$$
which, by permutations, is equivalent to computing the SVDs of the 2×2 matrices $\begin{pmatrix} 1 & -2\sigma_j \\ 0 & -1 \end{pmatrix}$. This is, in turn, a simple calculation. ∎

We note that for the solution R of the Sylvester equation (12) we have
$$\|R\| \leq \frac{\|A_{12}\|}{\operatorname{sep}(A_{11}, A_{22})},$$
where the equality is attainable [33]. From Lemma 3.1, we see that the conditioning of the matrix sign function computation is closely related to the norm of the projection P, therefore the norm of R, which in turn is closely related to the quantity $\operatorname{sep}(A_{11}, A_{22})$. Specifically, when $\|R\|$ is large,
$$\|S\| = \|\operatorname{sign}(A)\| \leq \frac{2\|A_{12}\|}{\operatorname{sep}(A_{11}, A_{22})} \tag{13}$$
and
$$\kappa(S) \leq \frac{4\|A_{12}\|^2}{\operatorname{sep}^2(A_{11}, A_{22})}.$$
If $\|A_{12}\|$ is moderate, an ill-conditioned matrix sign function means large $\|R\|$, which in turn means small $\operatorname{sep}(A_{11}, A_{22})$. Following Stewart [33], it means that it is harder to separate the invariant subspaces corresponding to the matrices A_{11} and A_{22}.

The following theorem discusses the conditioning of the eigenvalues of $\operatorname{sign}(A)$ and the distance from $\operatorname{sign}(A)$ to the nearest ill-posed problem.

THEOREM 3.2. *Let A and R be as in Lemma* 3.1. *Then we have the following:*

1. *Let δS have the property that $S + \delta S$ has a pure-imaginary eigenvalue. Then δS may be chosen with $\|\delta S\| = 1/\|S\|$ but no smaller. In the language of* [35], *the ϵ-pseudospectrum of S excludes the imaginary axis for $\epsilon < 1/\|S\|$, and intersects it for $\epsilon \geq 1/\|S\|$.*
2. *The condition number of the eigenvalues of S is $\|P\|$. In other words, perturbing S by a small δS perturbs the eigenvalues by at most $\|P\|\ \|\delta S\| + \mathcal{O}(\|\delta S\|^2)$.*
3. *If A is close to S and $\kappa(S) < \mathbf{u}^{-1/2}$, then Newton iteration* (2) *in floating point arithmetic will compute S with an absolute error bounded by $\mathbf{u}^{1/2}\|S\|$.*

Proof.

1. The problem is to minimize $\sigma_{\min}(S - i\zeta I)$ over all real ζ, where $\sigma_{\min}$ is the smallest singular value of $S - i\zeta I$. Using the same unitary similarity transformation and permutation as in the part 1 of Lemma 3.1, we see that this is equivalent to minimizing
$$\sigma_{\min}\left(\begin{pmatrix} 1-i\zeta & -2\sigma_j \\ 0 & -1-i\zeta \end{pmatrix}\right)$$
over all σ_j and real ζ. This is a straightforward calculation, with the minimum being obtained for $\zeta = 0$ and $\sigma_j = \|R\|$.
2. The condition number of a semisimple eigenvalue is equal to the secant of the acute angle between its left and right eigenvectors [24, 17]. Using the above reduction to 2×2 subproblems (this unitary transformation of coordinates does not changes angles between vectors), this is again a straightforward calculation.
3. Since $\|S\| = \|S^{-1}\|$, the absolute error δS in computing $\frac{1}{2}(S+S^{-1})$ is bounded essentially by the error in computing S^{-1}:
$$\|\delta S\| \lesssim \mathbf{u}(\|S\| \cdot \|S^{-1}\|)\|S^{-1}\| = \mathbf{u}\|S\|^3 < \mathbf{u}^{1/2}\|S\|.$$
For the Newton iteration to converge, δS cannot be so large that $S + \delta S$ has pure-imaginary eigenvalues; from the result 1, this means $\|\delta S\| < \|S\|^{-1}$. Therefore, if $\mathbf{u}^{1/2}\|S\| < \|S\|^{-1}$, i.e., $\kappa(S) < \mathbf{u}^{-1/2}$, then Newton iteration (2) will compute S with an absolute error bounded by $\mathbf{u}^{1/2}\|S\|$. ∎

It is naturally desired to have an analysis from which we know the conditioning of the intermediate matrices A_k in the Newton iteration. It will help us in addressing the question of how to detect the possible appearance of pure-imaginary eigenvalues and to modify or terminate the iteration early if necessary. Unfortunately, it is difficult to make a clean analysis far from convergence because we are unable to relate the error in each step of the iteration to the conditioning of the problem. We can do a coarse analysis, however, in the case that the matrix is diagonalizable.

THEOREM 3.3. *Let A have eigenvalues λ_j (none pure imaginary or zero), right eigenvectors x_j, and left eigenvectors y_j normalized so $\|x_j\| = \|y_j\| = 1$. Let $s_j = \operatorname{sign}(\Re(\lambda_j))$, and*
$$\sigma = \min_j \frac{|y_j^H x_j|}{n} \cdot \frac{|\lambda_j + s_j| - |\lambda_j - s_j|}{|\lambda_j + s_j| + |\lambda_j - s_j|}. \tag{14}$$
Let A_k be the matrix obtained at the kth Newton iteration (2). Then for all k, $\sigma_{\max}(A_k) \le 1/\sigma$ and $\sigma_{\min}(A_k) \ge \sigma$, i.e.,
$$\kappa(A_k) = \frac{\sigma_{\max}(A_k)}{\sigma_{\min}(A_k)} \le \frac{1}{\sigma^2}. \tag{15}$$

Proof. We may express the eigendecomposition of A as $A = \sum_{j=1}^n \lambda_j x_j y_j^H / y_j^H x_j$. Then $A_k = \sum_{j=1}^n \lambda_{j,k} x_j y_j^H / y_j^H x_j$, where $\lambda_{j,k} = \frac{1}{2}(\lambda_{j,k-1}^{-1} + \lambda_{j,k-1})$ with $\lambda_{j,0} = \lambda_j$. We wish to bound $|\lambda_{j,k}|$ from above and below for all k. This is easily done by noting that
$$\frac{\lambda_{j,k+1} - s_j}{\lambda_{j,k+1} + s_j} = \left(\frac{\lambda_{j,k} - s_j}{\lambda_{j,k} + s_j}\right)^2$$

so that all $\lambda_{j,k}$ lie inside a disk defined by

$$\left|\frac{\lambda_{j,k}-s_j}{\lambda_{j,k}+s_j}\right| \le \left|\frac{\lambda_j-s_j}{\lambda_j+s_j}\right| \equiv c_j < 1.$$

This disk is symmetric about the real axis, so its points of minimum and maximum absolute value are both real. Solving for these extreme points yields

$$\frac{1-c_j}{1+c_j} \le |\lambda_{j,k}| \le \frac{1+c_j}{1-c_j}.$$

This means

$$\sigma_{\max}(A_k) = \|A_k\| = \left\|\sum_{j=1}^{n} \lambda_{j,k}\frac{x_j y_j^H}{y_j^H x_j}\right\| \le \sum_{j=1}^{n}\frac{|\lambda_{j,k}|}{|y_j^H x_j|} \le \max_j \frac{n}{|y_j^H x_j|}\cdot\frac{1+c_j}{1-c_j}.$$

Similarly

$$\sigma_{\min}^{-1}(A_k) = \|A_k^{-1}\| = \left\|\sum_{j=1}^{n} \lambda_{j,k}^{-1}\frac{x_j y_j^H}{y_j^H x_j}\right\| \le \sum_{j=1}^{n}\frac{|\lambda_{j,k}^{-1}|}{|y_j^H x_j|} \le \max_j \frac{n}{|y_j^H x_j|}\cdot\frac{1+c_j}{1-c_j},$$

which proves the bound (15). □

As we know, the error introduced at each step of the iteration is mainly caused by the computation of matrix inverse, which is approximately bounded in norm by

$$\mathbf{u}(\kappa(A_k)\|A_k^{-1}\| + \|A_k\|) \le \mathbf{u}(\sigma^{-3}+\sigma^{-1}) \approx \mathbf{u}\sigma^{-3}$$

when $\sigma \ll 1$. If $\mathbf{u}\sigma^{-3} < \sigma_{\min}(A_k)$, then this error cannot make an intermediate A_k become singular and cause the iteration to fail. Our analysis shows that if $\mathbf{u}\sigma^{-3} < \sigma$, or $\sigma > \mathbf{u}^{1/4}$, then the iteration will not fail. This very coarse bound generalizes result 3 of Theorem 2.

We note that if A is symmetric, by the orthonormal eigendecomposition of $A = \sum_{j=1}^{n}\lambda_j q_j q_j^T$, where $q_j^T q_j = 1$, $q_j^T q_k = 0$ if $j \ne k$, then from Theorem 3 we have

$$\sigma = \min_j \begin{cases} \frac{1}{|\lambda_j|} & \text{if } |\lambda_j| \ge 1, \\ |\lambda_j| & \text{if } |\lambda_j| < 1. \end{cases}$$

Therefore,

$$\kappa(A_k) \le \max_j \begin{cases} \lambda_j^2 & \text{if } |\lambda_j| \ge 1, \\ \frac{1}{\lambda_j^2} & \text{if } |\lambda_j| < 1. \end{cases} \tag{16}$$

It shows that when A is symmetric, the condition number of the intermediate matrices A_k, which affects the numerical stability of the Newton iteration, is essentially determined by the square of the distance of the eigenvalues to the imaginary axis.[1]

When A is nonsymmetric and diagonalizable, from Theorem 3.3, we also see that the condition number of the intermediate matrices A_k is related to the norms

[1] A referee predicted that in the symmetric case, the condition number of A_k might be determined only by the distance, not the square of the distance. We were not able to prove such prediction.

of the spectral projectors $P_j = x_j y_j^H/(y_j^H x_j)$ corresponding to the eigenvalues λ_j ($\|P_j\| = 1/|y_j^H x_j|$) and the quantities of the form

$$\tilde{\sigma}_j = \frac{|\lambda_j + s_j| - |\lambda_j - s_j|}{|\lambda_j + s_j| + |\lambda_j - s_j|},$$

where $s_j = \text{sign}(\Re(\lambda_j))$. If we write $\lambda_j = \alpha_j + i\beta_j$, by a simple algebraic manipulation, we have

$$\tilde{\sigma}_j = \frac{1}{2|\alpha_j|}\left[1 + \alpha_j^2 + \beta_j^2 - \sqrt{(\alpha_j^2 - 1)^2 + 2(1 + \alpha_j^2)\beta_j^2 + \beta_j^4}\,\right].$$

From this expression, we see that if there is an eigenvalue λ_j of A very near to the pure imaginary axis, i.e., α_j is small, then by the first-order Taylor expansion of $\tilde{\sigma}_j$ in terms of α_j, we have

$$\tilde{\sigma}_j = \frac{|\alpha_j|}{1 + \beta_j^2} + \mathcal{O}(\alpha_j^2). \tag{17}$$

Therefore, to first order in α_j, the condition numbers of the intermediate matrices A_k satisfy

$$\kappa(A_k) \leq \frac{1}{\sigma^2} = \max_j \left(\frac{|\alpha_j|}{n\|P_j\|(1 + \beta_j^2)} + \mathcal{O}\left(\frac{\alpha_j^2}{\|P_j\|}\right)\right)^{-2}. \tag{18}$$

This implies that even if the eigenvalues of A are well conditioned (i.e., the $\|P_j\|$ are not too large), if there are also eigenvalues of A closer to the imaginary axis than $\mathbf{u}^{1/2}$, then the condition number of A_k could be large, $\kappa(A_k) \geq \mathbf{u}^{-1}$, and so the Newton iteration could fail to converge.

4. Backward stability of computed invariant subspace. As discussed in the previous section, because of possible ill conditioning of a matrix with respect to inversion and rounding errors during the Newton iteration, we generally only expect to be able to compute the matrix sign function to the square root of the machine precision, provided that the initial matrix A has condition number smaller than $\mathbf{u}^{-1/2}$. This means that when Newton iteration converges, the computed matrix sign function $\widehat{S}$ satisfies

$$\widehat{S} = S + F \quad \text{with} \quad \|F\| \leq \mathcal{O}(\sqrt{\mathbf{u}})\|S\|. \tag{19}$$

Under this assumption, $\widehat{P} = \frac{1}{2}(\widehat{S} + I)$ is an approximate spectral projection corresponding to $\lambda_+(A)$. Therefore, if $\ell = \text{rank}(\widehat{P})$, the first ℓ columns $\widehat{Q}_1$ of $\widehat{Q} \equiv Q + \delta Q = (\widehat{Q}_1,\ \widehat{Q}_2)$ in the rank revealing QR decomposition of $\widehat{P}$ span an approximate invariant subspace. $\widehat{Q}^H A \widehat{Q}$ has the form

$$\widehat{Q}^H A \widehat{Q} = (Q + \delta Q)^H A (Q + \delta Q) = \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ E_{21} & \widehat{A}_{22} \end{pmatrix}$$

with $\lambda(\widehat{A}_{11})$ being the approximate eigenvalues of A in $\mathbf{C}_+$, and $\lambda(\widehat{A}_{22})$ being the approximate eigenvalues of A in $\mathbf{C}_-$. Since we expect the computed matrix sign function to be of half-machine precision, it is reasonable to expect computing the

invariant subspace to half-precision too. This in turn means that the backward error $\|E_{21}\|$ in the computed decomposition $\widehat{Q}^H A\widehat{Q}$ is bounded by $\mathcal{O}(\sqrt{\mathbf{u}})\|A\|$, provided that the problem is not very ill conditioned. In this section, we will try to justify such expectation.

To this end, we first need to bound the error in the space spanned by the leading $\ell = \mathrm{rank}(P)$ columns of the transformation matrix Q, i.e., we need to know how much a right singular subspace of the exact projection matrix $P = \frac{1}{2}(S + I)$ is perturbed when P is perturbed by a matrix of norm η. Since P is a projector, the subspace is spanned by the right singular vectors corresponding to all nonzero singular values of P (call the set of these singular values $\mathcal{S}$). In practice, of course, this is a question of rank determination. From the well-known perturbation theory of the singular value decomposition [34, page 260], the space spanned by the corresponding singular vectors is perturbed by at most $\mathcal{O}(\eta)/\mathrm{gap}_{\mathcal{S}}$, where $\mathrm{gap}_{\mathcal{S}}$ is defined by

$$\mathrm{gap}_{\mathcal{S}} \equiv \min_{\substack{\sigma \in \mathcal{S} \\ \bar{\sigma} \notin \mathcal{S}}} |\sigma - \bar{\sigma}| \ .$$

To compute $\mathrm{gap}_{\mathcal{S}}$, we note that there is always a unitary change of basis in which a projector is of the form $\begin{pmatrix} I & \Sigma \\ 0 & 0 \end{pmatrix}$, where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_l)$ is diagonal with $\sigma_1 \geq \cdots \geq \sigma_l \geq 0$. By straightforward calculation, we find that the singular values of the projector are $\{\sqrt{1+\sigma_1^2}, \ldots, \sqrt{1+\sigma_\ell^2}, 1, \ldots, 1, 0, \ldots, 0\}$, where the number of ones in the set of singular values is equal to $\max\{2\ell - n, 0\}$. Since $\mathcal{S} = \{\sqrt{1+\sigma_1^2}, \ldots, \sqrt{1+\sigma_\ell^2}, 1, \ldots, 1\}$, we have

$$\mathrm{gap}_{\mathcal{S}} = \begin{cases} \sqrt{1+\sigma_\ell^2} & \text{if } 2\ell \leq n, \\ 1 & \text{if } 2\ell > n. \end{cases}$$

Thus, the error δQ in Q is bounded by

$$\|\delta Q\| \leq \frac{\mathcal{O}(\|F\|)}{\mathrm{gap}_{\mathcal{S}}} \leq \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\mathrm{gap}_{\mathcal{S}}}. \tag{20}$$

Hence, the backward error in the computed spectral decomposition is bounded by

$$\begin{aligned} \|E_{21}\| &\leq \|(Q+\delta Q)^H A(Q+\delta Q) - Q^H A Q\| \\ &= \|\delta Q^H A Q + Q^H A \delta Q + \delta Q^H A \delta Q\| \\ &\leq 2\|\delta Q\| \, \|A\| + \mathcal{O}(\epsilon^2), \end{aligned}$$

where $\mathcal{O}(\epsilon^2)$ is the second-order perturbation term of $\|\delta Q\|$. Therefore, if $2\ell \leq n$, we have the following first-order bound on the backward stability of computed invariant subspace:

$$\frac{\|E_{21}\|}{\|A\|} \leq \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\mathrm{gap}_{\mathcal{S}}} = \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|S\|}{\sqrt{1+\sigma_\ell^2}}. \tag{21}$$

If we use the bound (5) of the matrix sign function S, then from (21) we have

$$\frac{\|E_{21}\|}{\|A\|} \leq \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|A\|}{d_A\sqrt{1+\sigma_\ell^2}}, \tag{22}$$

where d_A, defined in (4), is the distance to the ill-posed problem. On the other hand, if we use the bound (13) for the matrix sign function S, then from (21) again we have

$$\frac{\|E_{21}\|}{\|A\|} \leq \frac{\mathcal{O}(\sqrt{\mathbf{u}})\|A\|}{\delta\sqrt{1+\sigma_\ell^2}}, \tag{23}$$

where $\delta = \mathrm{sep}(A_{11}, A_{22})$ is the separation of the matrices A_{11} and A_{22}, if A is assumed to have the form (11). We note that the error bound (23) is essentially the same as the error bound given by Byers, He, and Mehrmann [13], although we use a different approach. In [13], it is assumed that $\|F_{21}\| \lesssim \mathcal{O}(\mathbf{u})\|S\|$ in (19), where F_{21} is the (2,1) block of the matrix F. Therefore, the $\mathcal{O}(\sqrt{\mathbf{u}})$ term in (23) is replaced by $\mathcal{O}(\mathbf{u})$.

The bounds (22) and (23) reveal two important features of the matrix-sign-function-based algorithm for computing the invariant subspace. First, they indicate that the backward error in the computed approximate invariant subspace appears no larger than the absolute error in the computed matrix sign function, provided that the spectral decomposition problem is not very ill conditioned (i.e., d_A or δ is not tiny). Second, if $2\ell \leq n$, the backward error is a decreasing function of σ_l. If σ_ℓ is large, this means σ_1 and so $\|P\| = \sqrt{1+\sigma_1^2}$ are large, and this in turn means the eigenvalues close to the imaginary axis are ill conditioned. It is harder to divide these eigenvalues. Of course as they become ill conditioned, d_A decreases at the same time, which must counterbalance the increase in σ_ℓ in a certain range.

It is interesting to ask which error bound (22) and (23) is sharper, i.e., which one of the quantities d_A and $\delta = \mathrm{sep}(A_{11}, A_{22})$ is larger. In [13], an example of a 2×2 matrix is given to show that the quantity δ is larger than the quantity d_A. However, we can also devise simple examples to show that d_A can be larger than $\delta = \mathrm{sep}(A_{11}, A_{22})$. For example, let $A = \mathrm{diag}(A_{11}, A_{22})$ with

$$A_{11} = \begin{pmatrix} \eta & 2 & 3 \\ 0 & \eta & 2 \\ 0 & 0 & \eta \end{pmatrix}, \qquad A_{22} = \begin{pmatrix} -\eta & 2 & 3 \\ 0 & -\eta & 2 \\ 0 & 0 & -\eta \end{pmatrix}.$$

When $\eta = 10^{-3}$, we have $d_A \approx 2.50 \times 10^{-10}$, and $\delta = \mathrm{sep}(A_{11}, A_{22}) \approx 2.81 \times 10^{-16}$. More generally, by choosing A_{11} to be a large Jordan block with a tiny eigenvalue, and $A_{22} = -A_{11}$, d_A is close to the square root of δ. d_A is computed using "numerical brute force" to plot the function $d_A(\tau)$ on a wide range of $\tau \in \mathbb{R}$, and search for the minimal value.

Note that by modifying A to be $A - \sigma I$, where σ is a (sufficiently small) real number, d_A will change but δ will not. Thus, d_A and δ are not completely comparable quantities. We believe d_A to be a more natural quantity to use than δ, since δ does not always depend on the distance to the nearest ill-posed problem. This is reminiscent of the difference between the quantities $\delta = \mathrm{sep}(A_{11}, A_{22})$ and $\mathrm{sep}_\lambda(A_{11}, A_{22})$ [18].

In practice, we will use the a posteriori bound $\|E_{21}\|/\|A\|$ anyway, since if we block upper triangularize $\widehat{Q}^H A \widehat{Q}$ by setting the $(2,1)$ block to zero, $\|E_{21}\|/\|A\|$ is precisely the backward error we introduce.

Before ending this section, let us comment on the stability of the matrix-sign-function-based algorithm versus the QR algorithm. The QR algorithm is a numerical backward stable method for computing the Schur decomposition of a general non-symmetric matrix A. The computed Schur form $\widehat{T}$ and Schur vectors $\widehat{Q}$ by the QR algorithm satisfy

$$\widehat{Q}^H(A+E)\widehat{Q} = \widehat{T},$$

where E is of the order of $\mathbf{u}\|A\|$. Numerical software for the QR algorithm is available in EISPACK [32] and LAPACK [1]. Although nonconvergent examples have been found, they are quite rare in practice [6, 16]. We note that the eigenvalues on the (block) diagonal of $\widehat{T}$ may appear in any order. Therefore, if an application requires an invariant subspace corresponding to the eigenvalues in a specific region in complex plane, a second step of reordering eigenvalues on the diagonal of $\widehat{T}$ is necessary. A guaranteed stable implementation of this reordering is described in [7].

The matrix-sign-function-based algorithm can be regarded as an algorithm to combine these two steps into one. If the matrix sign function can be computed within the order of $\mathbf{u}\|S\|$, then the analysis in this section shows that the matrix-sign-function-based algorithm could be as stable as the QR algorithm plus reordering. Unfortunately, if the matrix is ill conditioned with respect to matrix inversion (which does not affect the QR algorithm), numerical instability is anticipated in the computed matrix sign function. Therefore, in general, the matrix sign function is less stable than the QR algorithm plus reordering.

5. Numerical experiments. In this section, we will present numerical examples to verify the above analysis. We will see the numerical stability of the Newton iteration (2) and the backward accuracy of computed spectral decomposition (1) under the influence of the conditioning of the matrix A with respect to inversion, the condition number $\kappa(S)$ of $S = \mathrm{sign}(A)$, and the distance $\Delta(A)$ of the eigenvalues of A to the pure-imaginary axis, where $\Delta(A) = \min_i |\Re(\lambda_i(A))|$. We use the easily computed quantity $\Delta(A)$ as a surrogate of the quantity d_A in (4).

Let us recall that the analysis of sections 3 and 4 essentially claims the following:

(1) If $\Delta(A) < \mathbf{u}^{1/2}$, then the Newton iteration may fail to converge or fail to compute the matrix sign function within the absolute error $\mathbf{u}^{1/2}\|S\|$, even when the matrix sign function is well conditioned. See (18).
(2) If $\kappa(S) > \mathbf{u}^{-1/2}$, then even the distance $\Delta(A)$ is not small, and the Newton iteration may still fail to compute the matrix sign function in the absolute error of $\mathcal{O}(\mathbf{u}^{1/2}\|S\|)$. See part 3 of Theorem 3.2.
(3) In general, the backward error in the computed spectral decomposition will be smaller than the absolute error in the computed matrix sign function. See (21).

The following numerical examples will illustrate these claims. Our numerical experiments were performed on a SUN workstation 10 with machine precision $\varepsilon_M = 2.2204 \times 10^{-16} \approx \mathbf{u}$. All the algorithms are implemented in MATLAB 4.0a. We use the simple Newtion iteration (2) to compute the matrix sign function with the stopping criterion

$$\|A_{k+1} - A_k\| \leq 10n\varepsilon_M\|A_k\|.$$

The maximal number of iterations is set to be 70. At the convergence, we have $\lim_{k\to\infty} A_k = \widehat{S}$, the computed matrix sign function. We use the QR decomposition with column pivoting as the rank revealing scheme. $\frac{1}{2}(\widehat{S} + I) = \widehat{Q}\widehat{R}\Pi$, and finally compute

$$\widehat{Q}^H A \widehat{Q} = \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ E_{21} & \widehat{A}_{22} \end{pmatrix},$$

where the first $\ell = \mathrm{rank}(\widehat{R})$ columns of $\widehat{Q}$ spans the invariant subspaces corresponding

TABLE 1
Numerical results for Example 1.

c	$\Delta(A) = s$	$\kappa(A)$	$\kappa(S)$	iter	$\frac{\|S-\bar{S}\|}{\|S\|}$	$\frac{\|E_{21}\|}{\|A\|}$
10	$1.0\text{e}+00$	$1.9\text{e}+03$	$2.7\text{e}+03$	7	$2.9\text{e}-14$	$3.9\text{e}-17$
	$1.0\text{e}-02$	$8.6\text{e}+02$	$1.5\text{e}+02$	13	$8.4\text{e}-14$	$8.4\text{e}-16$
	$1.0\text{e}-04$	$3.8\text{e}+01$	$8.1\text{e}+01$	20	$1.3\text{e}-11$	$1.3\text{e}-13$
	$1.0\text{e}-06$	$4.7\text{e}+02$	$9.0\text{e}+02$	30	$4.1\text{e}-09$	$4.1\text{e}-12$
	$1.0\text{e}-08$	$4.3\text{e}+02$	$1.0\text{e}+03$	33	$2.8\text{e}-07$	$2.8\text{e}-10$
	$1.0\text{e}-09$	$2.8\text{e}+02$	$1.8\text{e}+03$	36	$8.0\text{e}-06$	$8.0\text{e}-09$
	$1.0\text{e}-10$	$3.6\text{e}+01$	$3.7\text{e}+02$	40	$2.2\text{e}-05$	$2.2\text{e}-07$
	$1.0\text{e}-12$	$5.5\text{e}+01$	$1.0\text{e}+03$	46	$4.0\text{e}-03$	$4.0\text{e}-06$
10^3	$1.0\text{e}-06$	$7.8\text{e}+06$	$1.7\text{e}+07$	$26(10^{-11})$	$2.1\text{e}-06$	$5.4\text{e}-12$
	$1.0\text{e}-08$	$1.7\text{e}+06$	$1.0\text{e}+07$	$33(10^{-11})$	$5.1\text{e}-04$	$1.8\text{e}-09$

to $\lambda(\widehat{A}_{11})$, which are the approximate eigenvalues of A in $\mathbf{C}_+$. $\|E_{21}\|/\|A\|$ is the backward error committed by the algorithm.

All our test matrices are constructed of the form

$$A = U^T \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix} U, \tag{24}$$

where U is an orthogonal matrix generated from the QR decomposition of a random matrix with normal distribution having mean 0.0 and variance 1.0. We will choose different submatrices A_{11}, A_{22}, and A_{12} so that the generated matrices A have different specific features in order to observe our theoretical results in practice.

The exact matrix sign function $S = \text{sign}(A)$ of A and the condition number of S are computed as described in Lemma 3.1. The condition number of A is computed by MATLAB function `cond`.

In the following tables, iter is the number of iterations of the Newton iteration. A number 10^α in parenthesis next to an iteration number iter indicates that the convergence of the Newton iteration was stationary about $\mathcal{O}(10^\alpha)$ from the iter$^{\text{th}}$ iteration forward, and failed to satisfy the stopping criterion even after the allowed maximal number of iterations.

We have experimented with numerous matrices with different pathological ill conditioning in terms of the distance to the pure-imaginary axis, the condition numbers of $\kappa(A)$ and $\kappa(S)$, and the different values of $\text{sep}(A_{11}, A_{22})$ and so on. Two selected examples presented here are typical of behaviors we observed.

Example 1. In this example, the matrices A are of the form (24) with

$$A_{11} = \begin{pmatrix} s & 1 \\ -1 & s \end{pmatrix}, \quad A_{22} = \begin{pmatrix} -s & 1 \\ -1 & -s \end{pmatrix},$$

and $A_{12} = -(A_{11}R - RA_{22})$, where R is a random 2×2 matrix with normal distribution $(0, 1)$ multiplying by a parameter c. The generated matrix A has two complex conjugate eigenpairs $s \pm i$ and $-s \pm i$. As $s \to 0$, the distance $\Delta(A) = s \to 0$ too. The size of the parameter c will adjust the conditioning of the resulted matrix A and its matrix sign function.

Table 1 reports the computed results for different values of $\Delta(A) = s$. From the table, we see that when the matrices are well conditioned and the corresponding

TABLE 2
Numerical results of Example 2.

d	$\Delta(A)$	$\kappa(A)$	$\kappa(S)$	iter	$\frac{\|S-\bar{S}\|}{\|S\|}$	$\frac{\|E_{21}\|}{\|A\|}$
1.0	1.2e − 01	5.8e + 02	6.4e + 01	9	1.4e − 14	1.7e − 15
0.7	6.5e − 02	1.3e + 04	1.3e + 04	10	2.0e − 13	1.6e − 14
0.5	8.5e − 02	3.4e + 04	2.2e + 05	$10(10^{-13})$	9.5e − 12	1.7e − 13
0.3	2.0e − 02	3.9e + 06	5.9e + 08	$10(10^{-09})$	3.5e − 09	4.6e − 11
0.25	6.2e − 03	2.9e + 07	1.2e + 09	$13(10^{-10})$	2.8e − 08	1.5e − 10
0.09	1.1e − 02	4.9e + 09	5.5e + 13	$12(10^{-06})$	3.0e − 03	1.0e − 07

matrix sign function is also well conditioned, as stated in the claim (1), the convergence rate and accuracy of the Newton iteration is clearly determined by the distance $\Delta(A)$. When the distance becomes smaller, there is a steady increase in the number of Newton iterations required to convergence and the loss of the accuracy in the computed matrix sign function and, therefore, the desired invariant subspace. From the table, we also see that when both $\Delta(A)$ and $\kappa(S)$ are moderate, the Newton iteration fails to compute the matrix sign function in half-machine precision. Nevertheless, the computed invariant subspace seems to still have half-machine precision; see the claim (3).

Example 2. In this example, the test matrices A are of the form (24). A_{12} are 5×5 $(1,0)$ normally distributed random matrices. The submatrices A_{11} and A_{22} are first set by 5×5 $(1,0)$ normally distributed random upper tridiagonal matrices, and then the diagonal elements of A_{11} and A_{22} are replaced by $d|a_{ii}|$ and $-d|a_{ii}|$, respectively, where $a_{ii}(1 \le i \le n)$ are random numbers with normal distribution $(0,1)$, d is a positive parameter. A_{12} are 5×5 $(1,0)$ normally distributed random matrices.

The numerical results are reported in Table 2. For the given parameter d, the eigenvalues are well separated away from the pure-imaginary axis ($\Delta(A)$ is not small), however, as stated in the claim (2), we see the influence of the condition numbers $\kappa(S)$ to the convergence of the Newton iteration and, therefore, the accuracy of the computed matrix sign function and the invariant subspace.

6. Refining estimates of approximate invariant subspaces. When we use the matrix-sign-function-based algorithm to deflate an invariant subspace of matrix A, we end up with the form

$$\widehat{Q}^H A\widehat{Q} = (\widehat{Q}_1,\ \widehat{Q}_2)^H A(\widehat{Q}_1,\ \widehat{Q}_2) = \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ E_{21} & \widehat{A}_{22} \end{pmatrix}, \tag{25}$$

where the size of $\|E_{21}\|/\|A\|$ reveals the accuracy and backward stability of computed invariant subspace spanning by $\widehat{Q}_1$ of A. If higher accuracy is desired, we may use iterative refinement techniques to improve the accuracy of computed invariant subspace. The methods are due to Stewart [33], Dongarra, Moler, and Wilkinson [20], and Chatelin [15]. Even though these methods all apparently solve different equations, as shown by Demmel [19], after changing variables, they all solve the same Riccati equation in the inner loop.

Let us follow Stewart's approach to present the first class of methods. From (25), we know that $\widehat{Q}_1$ spans an approximate invariant subspace and $\widehat{Q}_2$ spans an orthogonal complementary subspace. If we let the true invariant subspace be represented by

$\widehat{Q}_1 + \widehat{Q}_2 Y$ and, therefore, its orthogonal complementary subspace as $\widehat{Q}_2 - \widehat{Q}_1 Y^H$, then Y is derived as follows: $\widehat{Q}_1 + \widehat{Q}_2 Y$ will be an invariant subspace if and only if the lower left block of

$$(\widehat{Q}_1 + \widehat{Q}_2 Y,\ \widehat{Q}_2 - \widehat{Q}_1 Y^H)^{-1} A (\widehat{Q}_1 + \widehat{Q}_2 Y,\ \widehat{Q}_2 - \widehat{Q}_1 Y^H)$$

is zero, i.e., if the lower left corner of

$$\begin{pmatrix} I & -Y^H \\ Y & I \end{pmatrix} \begin{pmatrix} \widehat{A}_{11} & \widehat{A}_{12} \\ E_{21} & \widehat{A}_{22} \end{pmatrix} \begin{pmatrix} I & Y^H \\ -Y & I \end{pmatrix}$$

is zero. Thus, Y must satisfy the equation

$$\widehat{A}_{22} Y - Y \widehat{A}_{11} = E_{21} - Y \widehat{A}_{12} Y,$$

which is the well-known algebraic Riccati equation. We may use the following two iterative methods to solve it:

1. the simple Newton iteration

$$\widehat{A}_{22} Y_k - Y_k \widehat{A}_{11} = E_{21} - Y_{k-1} \widehat{A}_{12} Y_{k-1} \tag{26}$$

   with $Y_0 = 0$, $k = 1, 2, \ldots$;
2. the modified Newton iteration

$$(\widehat{A}_{22} - Y_{k-1}\widehat{A}_{12}) Y_k - Y_k (\widehat{A}_{11} + \widehat{A}_{12} Y_{k-1}) = -E_{21} - Y_{k-1} \widehat{A}_{12} Y_{k-1} \tag{27}$$

   with $Y_0 = 0$, $k = 1, 2, \ldots$.

Therefore, we only need to solve a Sylvester equation in the inner loop of the iterative refinement.

In the following numerical example, we only use the simple Newton iteration (26) to refine the approximate invariant subspace computed by the matrix-sign-function-based algorithm, with the following stopping criterion:

$$\|Y_k - Y_{k-1}\|_1 \le 10 n \varepsilon_M \|Y_{k-1}\|_1.$$

Example 3. We continue Example 2. Table 3 lists the $\mathrm{sep}(A_{11}, A_{22})$, the number of iterative refinement steps, and the backward accuracy of improved invariant subspace.

As shown in the convergence analysis for the iterative solvers (26) and (27) of the Riccati equation by Stewart [33] and Demmel [18], if we let

$$\kappa = (\|\widehat{A}_{12}\|_F \|E_{21}\|_F)/\mathrm{sep}^2(\widehat{A}_{11}, \widehat{A}_{22}),$$

then under the assumptions $k < 1/4$ and $k < 1/12$, the iterations (26) and (27) converge, respectively. Therefore, $\mathrm{sep}(\widehat{A}_{11}, \widehat{A}_{22})$ is a key factor to the convergence of the iterative refinement schemes. The above examples verify such analysis. From the analysis of section 3, we recall that $\mathrm{sep}(\widehat{A}_{11}, \widehat{A}_{22})$ also affects the backward stability of the computed invariant subspace by the matrix-sign-function-based algorithm in the first place (before iterative refinement).

TABLE 3
Iterative refinement results of Example 2.

d	$\mathrm{sep}(A_{11}, A_{22})$	iter	$\frac{\|E'_{21}\|}{\|A\|}$
1.0	2.4e − 2	2	6.6e − 31
0.7	2.4e − 3	3	6.3e − 30
0.5	2.3e − 3	3	1.1e − 28
0.3	2.0e − 5	4	2.0e − 25
0.25	3.8e − 5	4	2.5e − 25
0.09	5.1e − 7	$5(10^{-12})$	1.1e − 21

7. Extension to the generalized eigenproblem. In this section, we outline a scheme to extend the matrix-sign-function-based algorithm to solve the generalized eigenvalue problem of a regular matrix pencil $A - \lambda B$. A matrix pencil $A - \lambda B$ is regular if $A-\lambda B$ is square and $\det(A-\lambda B)$ is not identically zero. In [22], Gardiner and Laub have considered an extension of the Newton iteration for computing the matrix sign function to a matrix pencil for solving generalized algebraic Riccati equations. Here we discuss another possible approach, which includes the computation of both left and right deflating subspaces.

For the given matrix pencil $A - \lambda B$, the problem of the spectral decomposition is to seek a pair of left and right deflating subspaces $\mathcal{L}$ and $\mathcal{R}$ corresponding to the eigenvalues of the pencil in a specified region $\mathcal{D}$ in complex plane. In other words, we want to find a pair of unitary matrices Q_L and Q_R so that if $Q_L = (Q_{L1}, Q_{L2})$, $\mathrm{span}(Q_{L1}) = \mathcal{L}$ and $Q_R = (Q_{R1}, Q_{R2})$, $\mathrm{span}(Q_{R1}) = \mathcal{R}$, then

$$Q_L^H A Q_R = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}, \quad Q_L^H B Q_R = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}, \tag{28}$$

where the eigenvalues of $A_{11} - \lambda B_{11}$ are the eigenvalues of $A - \lambda B$ in a selected region $\mathcal{D}$ in complex plane. Here, we will only discuss the region $\mathcal{D}$ to be the open *right* half-complex plane. As the same treatment in the standard eigenproblem, by employing Möbius transformations $(\alpha A + \beta B)(\gamma A + \delta B)^{-1}$ and divide-and-conquer, $\mathcal{D}$ can be the union of intersections of arbitrary half-planes and (complemented) disks, and so a rather general region.

To this end, by directly applying the Newton iteration to AB^{-1}, we have

$$Y_{k+1} = \frac{1}{2}(Y_k + Y_k^{-1}), \quad k = 0, 1, 2, \ldots, \quad Y_0 = AB^{-1}.$$

At convergence, $Y_\infty = \mathrm{sign}(AB^{-1})$. In practice, we do not want to invert B if it is ill conditioned. Hence, by letting $Z_k = Y_k B$, then the above iteration becomes

$$Z_{k+1}B^{-1} = \frac{1}{2}(Z_k B^{-1} + BZ_k^{-1}) = \frac{1}{2}(Z_k + BZ_k^{-1}B)B^{-1}.$$

This leads to the following iteration:

$$Z_{k+1} = \frac{1}{2}(Z_k + BZ_k^{-1}B)$$

for $k = 0, 1, 2, \ldots$ with $Z_0 = A$. Z_j converges quadratically to a matrix Z_∞. Then $Z_\infty B^{-1} = Y_\infty = \mathrm{sign}(AB^{-1})$. Next, to find the desired deflating subspace, we use

the rank revealing QR decomposition to calculate the range space of the projection $P = \frac{1}{2}(I + Z_\infty B^{-1})$ corresponding to the spectral in the open *right* half-plane, which has the same range space as $2PB = Z_\infty + B$. Thus, by computing the rank revealing QR decomposition of $Z_\infty + B = Q_L R_L \Pi_L$, we obtain the invariant subspace of AB^{-1} without inverting B, i.e.,

$$Q_L^H A B^{-1} Q_L = \begin{pmatrix} C_R & C_{12} \\ 0 & C_L \end{pmatrix}, \tag{29}$$

where $\lambda(C_R)$ are the eigenvalues of the pencil $A - \lambda B$ in the open right half-plane, $\lambda(C_L)$ are the ones of $A - \lambda B$ in the open left half-plane. Therefore, we have obtained the left deflating subspace of $A - \lambda B$.

To compute the right deflating subspace of $A - \lambda B$, we can apply the above idea to $A^H - \lambda B^H$, since transposing swaps right and left spaces. The Newton iteration implicitly applying to $A^H B^{-H}$ turns out to be

$$\tilde{Z}_{k+1} = \frac{1}{2}(\tilde{Z}_k + B^H \tilde{Z}_k^{-1} B^H)$$

for $k = 0, 1, 2, \ldots$ with $Z_0 = A^H$. $\tilde{Z}_j$ converges quadratically to a matrix $\tilde{Z}_\infty$. Using the same arguments as above, after computing the rank revealing QR decomposition of $\tilde{Z}_\infty - B = \tilde{Q}_R R_R \Pi_R$, we have

$$\tilde{Q}_R^H A^H B^{-H} \tilde{Q}_R = \begin{pmatrix} D_L & D_{12} \\ 0 & D_R \end{pmatrix},$$

where $\lambda(D_L)$ are the eigenvalues of the pencil $A - \lambda B$ in the open left half-plane, $\lambda(D_R)$ are the ones of $A - \lambda B$ in the open right half-plane. Note that for the desired spectral decomposition, after transposing, we need to first compute the deflating subspace corresponding to the eigenvalues in the open *left* half-plane. Let $Q_R = \tilde{Q}_R \tilde{\Pi}$, where $\tilde{\Pi}$ is an antidiagonal identity matrix[2]; then we have

$$Q_R^H A^H B^{-H} Q_R = \begin{pmatrix} D_R & 0 \\ D_{12} & D_L \end{pmatrix}. \tag{30}$$

From (29) and (30), we immediately have

$$Q_L^H A Q_R = \begin{pmatrix} C_R & C_{12} \\ 0 & C_L \end{pmatrix} Q_L^H B Q_R, \tag{31}$$

$$Q_L^H A Q_R = Q_L^H B Q_R \begin{pmatrix} D_R^H & D_{12}^H \\ 0 & D_L^H \end{pmatrix}. \tag{32}$$

Let $Q_L^H A Q_R$ and $Q_L^H B Q_R$ have the partitions

$$Q_L^H A Q_R = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad Q_L^H B Q_R = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix};$$

we have

$$\begin{pmatrix} C_R & C_{12} \\ 0 & C_L \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \begin{pmatrix} D_R^H & D_{12}^H \\ 0 & D_L^H \end{pmatrix}.$$

[2] The permutation $\tilde{\Pi}$ can be avoided if we use the rank revealing QL decomposition.

Then B_{21} satisfies

$$C_L B_{21} - B_{21} D_R^H = 0.$$

Note that $\lambda(C_L)$ are the eigenvalues of the pencil $A - \lambda B$ in the open left half-plane, $\lambda(D_R)$ are the eigenvalues of the pencil $A - \lambda B$ in the open right half-plane. Therefore, the above homogeneous Sylvester equation has only the solution $B_{21} = 0$. From (31) or (32), we have $A_{21} = 0$. The computed unitary orthogonal matrices Q_L and Q_R give the desired spectral decomposition (28).

8. Closing remarks. In this paper, we have presented a number of new results and approaches to further analyze the numerical behavior of the matrix sign function and algorithms using it to compute spectral decompositions of nonsymmetric matrices. From this analysis and numerical experiments, we conclude that if the spectral decomposition problem is not ill conditioned, the algorithm is a practical approach to solve the nonsymmetric eigenvalue problem. Performance evaluation of the matrix-sign-function-based algorithm on parallel distributed memory machines, such as the Intel Delta and CM-5, is reported in [4].

During the course of this work, we have discovered a new approach which essentially computes the same spectral projection matrix as the matrix sign function approach does, and also uses basic matrix operations, namely, matrix multiplication and the QR decomposition. However, it avoids the matrix inverse. From the point of view of accuracy, this is a more promising approach. The new approach is based on the work of Bulgakov and Godunov [10] and Malyshev [27, 28]. In [5], we have improved their results in several important ways, and made it a truly practical and *inverse-free* highly parallel algorithm for both the standard and generalized spectral decomposition problems. In brief, the difference between the matrix sign function and inverse-free methods is as follows. The matrix sign function method is significantly faster than the inverse-free method when it converges, but there are some very difficult problems where the inverse-free algorithm gives a more accurate answer than the matrix sign function algorithm. The interested reader may see paper [5] for details.

Acknowledgments. The authors would like to acknowledge Ralph Byers, Chunyang He, Nick Higham, and Volker Mehrmann for fruitful discussions on the subject. We would also like to thank the referees for their valuable comments on the manuscript.

The information presented here does not necessarily reflect the position or the policy of the U.S. Government and no official endorsement should be inferred.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, 2nd ed., SIAM, Philadelphia, PA, 1995.

[2] Z. Bai and J. Demmel, *Design of a parallel nonsymmetric eigenroutine toolbox, Part* I, in Proc. Sixth SIAM Conference on Parallel Processing for Scientific Computing, R. F. Sincovec et al., eds., SIAM, Philadelphia, PA, 1993; also available as Computer Science Report CSD-92-718, University of California, Berkeley, CA, 1992.

[3] Z. Bai and J. Demmel, *Design of a Parallel Nonsymmetric Eigenroutine Toolbox, Part* II, Department of Mathematics Research Report 95-11, University of Kentucky, Lexington, KY, 1995.

[4] Z. Bai, J. Demmel, J. Dongarra, A. Petitet, H. Robinson, and K. Stanley, *The spectral decomposition of nonsymmetric matrices on distributed memory parallel computers*, SIAM J. Sci. Comput., 18 (1997), pp. 1446–1461.

[5] Z. Bai, J. Demmel, and M. Gu, *Inverse free parallel spectral divide and conquer algorithms for nonsymmetric eigenproblems*, Numer. Math., 76 (1997), pp. 279–308.
[6] S. Batterson, *Convergence of the shifted QR algorithm on* 3 *by* 3 *normal matrices*, Numer. Math., 58 (1990), pp. 341–352.
[7] A. W. Bojanczyk and P. Van Dooren, *Reordering diagonal blocks in real Schur form*, in Linear Algebra for Large Scale and Real-Time Applications, G. H. Golub, M. S. Moonen, and B. L. R. De Moor, eds., Kluwer Academic Publishers, Amsterdam, 1993.
[8] S. Boyd and V. Balakrishnan, *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its* L_∞*-norm*, Systems Control Lett., 15 (1990), pp. 1–7.
[9] S. Boyd, V. Balakrishnan, and P. Kabamba, *A bisection method for computing the* H_∞ *norm of a transfer matrix and related problems*, Math. Control Signals Systems, 2 (1989), pp. 207–219.
[10] A. Y. Bulgakov and S. K. Godunov, *Circular dichotomy of the spectrum of a matrix*, Siberian Math. J., 29 (1988), pp. 734–744.
[11] R. Byers, *Solving the algebraic Riccati equation with the matrix sign function*, Linear Algebra Appl., 85 (1987), pp. 267–279.
[12] R. Byers, *A bisection method for measuring the distance of a stable matrix to the unstable matrices*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 875–881.
[13] R. Byers, C. He, and V. Mehrmann, *The Matrix Sign Function Method and the Computation of Invariant Subspaces*, Technical report preprint SPC 94-25, Fakultät für Mathematik, TU Chemnitz-Zwickau, Germany, 1994.
[14] R. Byers and N. K. Nichols, *On the stability radius of a generalized state-space system*, Linear Algebra Appl., 188–189 (1993), pp. 113–134.
[15] F. Chatelin, *Simultaneous Newton's iteration for the eigenproblem*, Comput. Suppl., 5 (1984), pp. 67–74.
[16] D. Day, *How the QR Algorithm Fails to Converge and How to Fix It*, Tech. Rep. SAND 96-09135, Sandia National Laboratories, Albuquerque, NM, 1996.
[17] J. Demmel, *The condition number of equivalence transformations that block diagonalize matrix pencils*, SIAM J. Numer. Anal., 20 (1983), pp. 599–610.
[18] J. Demmel, *On condition numbers and the distance to the nearest ill-posed problem*, Numer. Math., 51 (1987), pp. 251–289.
[19] J. Demmel, *Three methods for refining estimates of invariant subspaces*, Computing, 38 (1987), pp. 43–57.
[20] J. Dongarra, C. Moler, and J. H. Wilkinson, *Improving the accuracy of computed eigenvalues and eigenvectors*, SIAM J. Numer. Anal., 20 (1984), pp. 46–58.
[21] B. S. Garbow, J. M. Boyle, J. J. Dongarra, and C. B. Moler, *Matrix Eigensystem Routines – EISPACK Guide Extension*, Lecture Notes in Comput. Sci. 51, Springer-Verlag, Berlin, 1977.
[22] J. Gardiner and A. Laub, *A generalization of the matrix-sign function solution for algebraic Riccati equations*, Internat. J. Control, 44 (1986), pp. 823–832.
[23] N. J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra Appl., 212/213 (1994), pp. 3–20.
[24] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1980.
[25] C. Kenney and A. Laub, *Polar decomposition and matrix sign function condition estimates*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 488–504.
[26] C. Kenney, A. Laub, and P. Papadopoulos, *Matrix sign function algorithms for Riccati equations*, in Proc. of IMA Conference on Control: Modelling, Computation, Information, Southend-On-Sea, IEEE Press, Piscataway, NJ, 1992, pp. 1–10.
[27] A. N. Malyshev, *Guaranteed accuracy in spectral problems of linear algebra, Parts* I *and* II, Siberian Adv. Math., 2 (1992), pp. 144–197, 153–204.
[28] A. N. Malyshev, *Parallel algorithm for solving some spectral problems of linear algebra*, Linear Algebra Appl., 188/189 (1993), pp. 489–520.
[29] R. Mathias, *Condition estimation for the matrix function via the Schur decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 565–578.
[30] J. Roberts, *Linear model reduction and solution of the algebraic Riccati equation*, Internat. J. Control, 32 (1980), pp. 677–687.
[31] L. Shieh, H. Dib, and R. Yates, *Separation of matrix eigenvalues and structural decomposition of large-scale systems*, IEEE Proceedings: Control Theory Appl., 133 (1986), pp. 90–96.
[32] B. T. Smith, J. M. Boyle, J. J. Dongarra, B. S. Garbow, Y. Ikebe, V. C. Klema, and C. B. Moler, *Matrix Eigensystem Routines – EISPACK Guide*, Lecture Notes in Comput. Sci. 6, Springer-Verlag, Berlin, 1976.

[33] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.

[34] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[35] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991, D. F. Griffiths and G. A. Watson, eds., Longman Sci. Tech. Publ., Harlow, UK, 1992, pp. 234–266.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 226–234, January 1998

015

THE M-MATRIX GROUP GENERALIZED INVERSE PROBLEM FOR WEIGHTED TREES*

STEPHEN J. KIRKLAND† AND MICHAEL NEUMANN‡

Abstract. We characterize all weighted trees whose Laplacian has a group inverse which is an M-matrix. Actually, only a very narrow set of weighted trees yields such Laplacians. Our investigation involves analyzing circumstances under which a certain Z-matrix, derived from the tree and whose order is one less than the number of vertices in the tree, is an M-matrix. Our work here is motivated by a recent paper of Styan and Subak-Sharpe [*Linear Algebra Appl.*, 250 (1997), pp. 349–370] and by an earlier question of Deutsch and Neumann [*J. Math. Anal. Appl.*, 102 (1984), pp. 1–29].

Key words. Laplacian matrix, impedance matrix, weighted tree

AMS subject classifications. Primary, 05C50; Secondary, 15A48

PII. S0895479896304927

1. Introduction.

Motivated by (i) a paper of Styan and Subak-Sharpe [9], who asked when the *impedance* matrix of a resistive electrical network, subject to both Kirchoff's current and voltage laws, is an M-matrix and hence an *admittance* matrix in its own right, and (ii) an earlier question of Deutsch and Neumann [5], who asked when the group inverse of a singular and irreducible M-matrix is again an M-matrix, we characterize here the set of all weighted trees whose Laplacian has a group inverse which is an M-matrix.

As is well known (see Campbell and Meyer [3], Styan and Subak-Sharpe [9], and references cited therein), a resistive electrical network subject to Kirchoff's laws results in a consistent linear system $i = Yv$, where Y is a symmetric positive semidefinite M-matrix with zero row and column sums, the so-called *admittance* matrix for the network. The Moore–Penrose generalized inverse of Y is then the *impedance* matrix for the network.

Continuing, it is known from the theory of generalized inverses (see Ben-Israel and Greville [1] and Campbell and Meyer [3]) that for a symmetric matrix A, its *group inverse*, which exists, and its *Moore–Penrose generalized inverse* coincide. We refer the reader to these texts for more background material on generalized inverses and to the book by Berman and Plemmons [2] for background material concerning nonnegative matrices and M-matrices. In fact, all terminology and notations used in this paper come from the books just mentioned.

An *undirected weighted graph* on n vertices is a graph $\mathcal{G}$ each of whose edges e has been labeled by a positive real number $w(e)$ which is called the *weight* of the edge e. Taking the vertices of $\mathcal{G}$ to be $1, 2, \ldots, n$, the *Laplacian matrix* of the weighted graph $\mathcal{G}$ is the $n \times n$ matrix $L = (\ell_{i,j})$ whose ith diagonal entry equals the sum of the weights of the edges incident with vertex i and whose (i, j)th off-diagonal entry equals zero if

*Received by the editors June 3, 1996; accepted for publication (in revised form) by G. Styan February 8, 1997.

http://www.siam.org/journals/simax/19-1/30492.html

†Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan, Canada S4S 0A2 (kirkland@max.cc.uregina.ca). The research of this author was supported in part by NSERC grant OGP0138251.

‡Department of Mathematics, University of Connecticut, Storrs, CT 06269–3009 (neumann@math.uconn.edu). The research of this author was supported in part by NSF grant DMS–9306357.

there is no edge joining vertices i and j and otherwise is given by $-w(e)$, where e is the edge joining vertices i and j.

Suppose now that $\mathcal{T}$ is a weighted tree on n vertices and recall that any two vertices i and j are joined by a unique path $\mathcal{P}_{i,j}$. We define the *inverse weighted distance from vertex i to vertex j* as the sum

$$\tilde{d}(i,j) \;=\; \sum_{e\in\mathcal{P}_{i,j}} \frac{1}{w(e)},$$

that is, $\tilde{d}(i,j)$ is the sum of the reciprocals of the weights of the edges on the path $\mathcal{P}_{i,j}$. We define $\tilde{d}_{i,i} = 0$ for all $i = 1,\ldots,n$. For any vertex i, we define the *inverse status of vertex i* as the sum

$$\tilde{d}_i \;=\; \sum_{u\in T} \tilde{d}(u,i).$$

In a sequence of papers Kirkland, Neumann, and Shader [6], [7], [8] investigate applications of the group inverse of the Laplacian, such as to the determination of the algebraic connectivity of the graph. In this paper we shall make use of several results from [7] to show that there is a very limited set of weighted trees which yield a Laplacian whose group inverse is an M-matrix. A result which is particularly relevant to our investigation here is Corollary 3.8 in [7] which states that *a weighting of $\mathcal{T}$ yields that $L^{\#}$ is an* M*–matrix if and only if for any adjacent vertices i* and *j, with the edge e between them, we have*

$$\tilde{d}_i + \tilde{d}_j \;\leq\; \frac{n}{w(e)} + \frac{1}{n}\sum_{k=1}^{n} \tilde{d}_k. \tag{1.1}$$

We shall show that the set of all trees whose corresponding Laplacian has a group inverse which is an M-matrix consists only of certain weighted stars, which can be found in any order, and of certain weighted paths on four vertices. To arrive at our results we reformulate the conditions in (1.1) to show that the group inverse of the Laplacian is an M-matrix if and only if a certain $(n-1)\times(n-1)$ Z-matrix derived from the tree is an M-matrix. We then show that this is only possible in the restricted instances just mentioned.

2. Main results. Let $\mathcal{T}$ be a tree on n vertices. If e is an edge in $\mathcal{T}$, then $T\setminus e$ denotes the graph obtained from $\mathcal{T}$ by removing e. If i is a vertex of $\mathcal{T}$, then we define $\beta_i(e)$ to be the set of vertices in the connected component of $\mathcal{T}\setminus e$ which *does not* contain vertex i.

Recall next that for an $n\times n$ matrix Q, the unique $n\times n$ matrix X, if it exists, which satisfies the matrix equation $QXQ = Q$, $XQX = X$, and $QX = XQ$ is called the *group* (*generalized*) *inverse* of Q. It is known (see, for example, Ben-Israel and Greville [1] or Meyer and Campbell [3]) that the group inverse of Q exists if and only if the Jordan blocks of Q corresponding to the eigenvalue zero, if any, are all 1×1. In what follows we shall denote, as is customary, the group inverse of Q by $Q^{\#}$.

We begin with a result in which we recast the conditions in (1.1).

THEOREM 2.1. *Let $\mathcal{T}$ be a tree on n vertices. Label the vertices $1,\ldots,n$ and label the edges $e_1,\ldots,e_{n-1}$ with vertex i incident with edge e_i, $1\leq i\leq n-1$. Then there is a weighting of $\mathcal{T}$ whose Laplacian L has the property that $L^{\#}$ is an M-matrix if and*

only if the $(n-1)\times(n-1)$ *matrix*

$$A = \begin{bmatrix} [n|\beta_1(e_1)| - |\beta_1(e_1)|^2] & -|\beta_1(e_2)|^2 & \cdots & -|\beta_1(e_{n-1})|^2 \\ -|\beta_2(e_1)|^2 & [n|\beta_2(e_2)| - |\beta_2(e_2)|^2] & \cdots & -|\beta_2(e_{n-1})|^2 \\ \vdots & \cdots & \ddots & \vdots \\ -|\beta_{n-1}(e_1)|^2 & \cdots & \cdots & \cdot \end{bmatrix}$$

(2.1)

is an M-matrix.

Proof. Let i and j be adjacent vertices joined by the edge e. Then on examining the contribution of any edge f to the summand on the left-hand side of (1.1), we find that

$$\begin{aligned} \tilde{d}_i + \tilde{d}_j &= \left(\sum_{f\neq e} \frac{|\beta_i(f)|}{w(f)}\right) + \frac{|\beta_i(e)|}{w(e)} + \left(\sum_{f\neq e} \frac{|\beta_j(f)|}{w(f)}\right) + \frac{|\beta_j(e)|}{w(e)} \\ &= 2\left(\sum_{f\neq e} \frac{|\beta_i(f)|}{w(f)}\right) + \frac{n}{w(e)}. \end{aligned}$$

Thus, (1.1) holds if and only if

$$2\sum_{f\neq e} \frac{|\beta_i(f)|}{w(f)} \leq \frac{1}{n}\sum_{i=1}^{n} \tilde{d}_k.$$

Now, according to [7, Theorem 3.5],

$$\sum_{i=1}^{n} \tilde{d}_k = 2\sum_{f\in T} \frac{|\beta_i(f)|(n-|\beta_i(f)|)}{w(f)},$$

from which we see that (1.1) holds if and only if

$$\sum_{f\neq e} \frac{|\beta_i(f)|}{w(f)} \leq \sum_{f\in T} \frac{|\beta_i(f)|}{w(f)} - \frac{1}{n}\sum_{f\in T} \frac{(|\beta_i(f)|)^2}{w(f)}$$

or, equivalently, if and only if

$$\sum_{f\in T} \frac{(|\beta_i(f)|)^2}{w(f)} \leq n\frac{|\beta_i(e)|}{w(e)}.$$

Consequently, $L^{\#}$ is an M-matrix if and only if for any vertex i incident with edge e we have

$$n\frac{|\beta_i(e)|}{w(e)} - \sum_{f\in T} \frac{(|\beta_i(f)|)^2}{w(f)} \geq 0.$$

Using our labeling, this gives the condition

$$n\frac{|\beta_m(e_m)|}{w(e_m)} - \sum_{k=1}^{n-1} \frac{|\beta_m(e_k)|}{w(e_k)} \geq 0, \qquad 1\leq m\leq n-1, \tag{2.2}$$

as being necessary and sufficient for $L^{\#}$ to be an M-matrix.

Finally, note that in terms of the matrix A in (2.1), condition (2.2) means that

$$\begin{bmatrix} [n|\beta_1(e_1)| - |\beta_1(e_1)|^2] & -|\beta_1(e_2)|^2 & \cdots & -|\beta_1(e_{n-1})|^2 \\ -|\beta_2(e_1)|^2 & [n|\beta_2(e_2)| - |\beta_2(e_2)|^2] & \cdots & -|\beta_2(e_{n-1})|^2 \\ \vdots & \cdots & \ddots & \vdots \\ -|\beta_{n-1}(e_1)|^2 & \cdots & \cdots & \cdot \end{bmatrix} \begin{bmatrix} \frac{1}{w(e_1)} \\ \vdots \\ \vdots \\ \frac{1}{w(e_{n-1})} \end{bmatrix}$$

$$\geq 0.$$

Thus, according to Berman and Plemmons [2, Chapter 6, Theorem 4.16], we can conclude that $\mathcal{T}$ admits a weighting so that $L^{\#}$ is an M-matrix if and only if there exists a positive vector x so that

$$\begin{bmatrix} [n|\beta_1(e_1)| - |\beta_1(e_1)|^2] & -|\beta_1(e_2)|^2 & \cdots & -|\beta_1(e_{n-1})|^2 \\ -|\beta_2(e_1)|^2 & [n|\beta_2(e_2)| - |\beta_2(e_2)|^2] & \cdots & -|\beta_2(e_{n-1})|^2 \\ \vdots & \cdots & \ddots & \vdots \\ -|\beta_{n-1}(e_1)|^2 & \cdots & \cdots & \cdot \end{bmatrix} x \geq 0$$

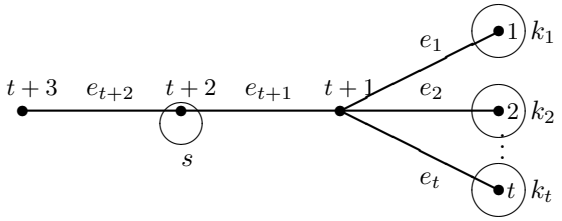
or, equivalently, because of irreducibility, if and only if the matrix in (2.1) is an M-matrix. □

In our next result, which is key to solving the M-matrix group inverse problem for weighted trees, we use Theorem 2.1 to determine which weighted trees, which are not stars, can possibly yield a Laplacian whose group inverse is an M-matrix. Note that if a tree is not a star, then necessarily it contains a path of length three as an induced subgraph.

THEOREM 2.2. *Suppose that $\mathcal{T}$ is a tree on n vertices which is not the star. Let u be a pendant vertex such that*

is a path of length three. Then the matrix in (2.1) *can be an M-matrix only if the set of vertices consisting of v and its branches which do not contain u or w has cardinality at least $(n-2)/2$.*

Proof. Since the tree under consideration has a path of length three with a vertex on this path being a pendant vertex, without loss of generality, we can assume that we have the following situation:



Here the circles on the right represent branches at $t+1$ containing vertices $1, \ldots, t$,

with the cardinality of the branch containing i being k_i, $i = 1, \ldots, t$. Similarly, s is the cardinality of the set of vertices consisting of vertex $t+2$ along with the branches at $t+2$ which contain neither $t+1$ nor $t+3$.

The proof will be done if we can show that

$$s \ \geq \ \frac{n-2}{2}.$$

We note that

$$\sum_{i=1}^{t} k_i \ = \ n - s - 2.$$

Moreover, $|\beta_i(e_i)| = n - k_i$ for all $1 \leq i \leq t$, $|\beta_i(e_j)| = k_j$ for all $i \neq j$ such that $1 \leq i, j \leq t$, $|\beta_i(e_{t+1})| = s+1$ for all $1 \leq i \leq t$, and $|\beta_i(e_{t+2})| = 1$ for all $1 \leq i \leq t$. Also,

$$|\beta_{t+1}(e_j)| \ = \ \begin{cases} k_j & \text{if } 1 \leq j \leq t, \\ s+1 & \text{if } j = t+1 \ , \\ 1 & \text{if } j = t+2 \end{cases}$$

and

$$|\beta_{t+2}(e_j)| \ = \ \begin{cases} k_j & \text{if } 1 \leq j \leq t, \\ n-1-s & \text{if } j = t+1 \ , \\ n-1 & \text{if } j = t+2. \end{cases}$$

It now follows that the principal submatrix of A determined by the indices $1, \ldots, t+2$ is

$$B \ = \ \left[\begin{array}{c|cc} nD - JD^2 & -(s+1)^2\mathbf{1} & -\mathbf{1} \\ \hline -\mathbf{1}^T D^2 & (s+1)(n-s-1) & -1 \\ -\mathbf{1}^T D^2 & -(n-1-s)^2 & (n-1) \end{array}\right],$$

where $D = \mathrm{diag}(k_1, \ldots, k_t)$, where $\mathbf{1}$ is the t-vector of all ones, and where $J = \mathbf{1}^T\mathbf{1}$. If A is an M-matrix, then, necessarily, so is B and, in particular according to a result of Crabtree [4], so is the Schur complement of B on its last two rows and columns. Note that $\mathbf{1}^T D\mathbf{1} = \sum_{i=1}^{t} k_i = n - s - 2$. Now,

$$\left(nD - JD^2\right)^{-1} \;=\; \frac{1}{n}D^{-1}\left(I + \frac{1}{s+2}JD\right),$$

where I is the $t \times t$ identity matrix, and we find that

$$\begin{aligned}
&\left[\begin{array}{c} \mathbf{1}^T D^2 \\ \hline \mathbf{1}^T D^2 \end{array}\right] \frac{1}{n}D^{-1}\left(I + \frac{1}{s+2}JD\right)\left[\begin{array}{c|c} (s+1)^2\mathbf{1} & \mathbf{1} \end{array}\right] \\
&\qquad = \frac{1}{n}\left[\begin{array}{c} \mathbf{1}^T D\left(1 + \dfrac{n-s-2}{s+2}\right) \\ \hline \mathbf{1}^T D\left(1 + \dfrac{n-s-2}{s+2}\right) \end{array}\right]\left[\begin{array}{c|c} (s+1)^2\mathbf{1} & \mathbf{1} \end{array}\right] \\
&\qquad = \frac{n-s-2}{s+2}\left[\begin{array}{cc} (s+1)^2 & 1 \\ (s+1)^2 & 1 \end{array}\right].
\end{aligned}$$

Hence, our Schur complement is given by

$$\begin{aligned}
&\left[\begin{array}{cc} (s+1)(n-s-1) & -1 \\ -(n-1-s)^2 & (n-1) \end{array}\right] - \frac{(n-s-2)}{s+2}\left[\begin{array}{cc} (s+1)^2 & 1 \\ (s+1)^2 & 1 \end{array}\right] \\
&\qquad = \frac{1}{s+2}\left[\begin{array}{cc} (s+1)n & -n \\ \left\{-(s+2)\left[n^2 - 2n(s+1)\right] - n(s+1)^2\right\} & n(s+1) \end{array}\right] \\
&\qquad = \frac{n}{s+2}\left[\begin{array}{cc} s+1 & -1 \\ -(s+2)\left[n - 2(s+1)\right] - (s+1)^2 & s+1 \end{array}\right].
\end{aligned}$$

But the last matrix displayed is an M-matrix if and only if

$$(s+1)^2 - (s+2)\left[n - 2(s+1)\right] - (s+1)^2 \;\geq\; 0,$$

that is, if and only if $2(s+1) - n \geq 0$ or, equivalently, if and only if $s \geq (n-2)/2$ which is the desired inequality. □

We can now apply the results of Theorem 2.2 to characterize all weighted trees whose Laplacian has a group inverse which is an M-matrix.

THEOREM 2.3. *Let $\mathcal{T}$ be a tree on n vertices. Then $\mathcal{T}$ admits a weighting such that $L^{\#}$ is an M-matrix if and only if either $n \leq 4$ or $n \geq 5$ and $\mathcal{T}$ is a star. The weightings for the n star which yield an M-matrix are of the form*

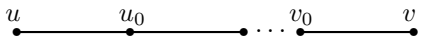
$$\left[\begin{array}{ccc} \frac{1}{w(e_1)} & \cdots & \frac{1}{w(e_{n-1})} \end{array}\right]^T \;=\; (I+J)y$$

for some nonzero nonnegative vector y. For $n = 4$, there is just one class of weightings of the path which yields an M-matrix given by

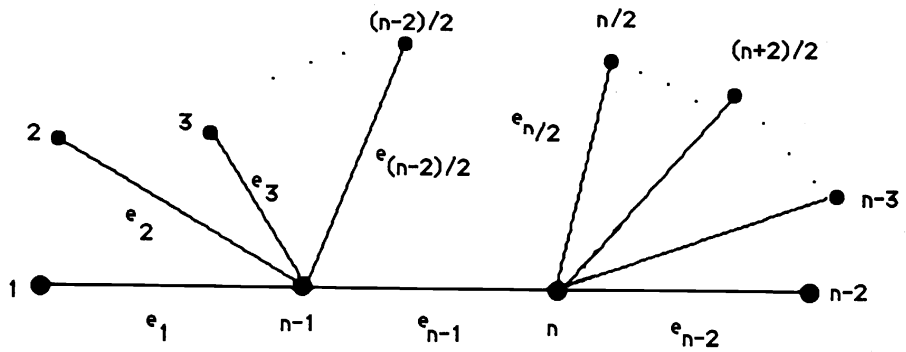
$$\left[\begin{array}{ccc} w(e_1) & w(e_2) & w(e_3) \end{array}\right] \;=\; \theta\left[\begin{array}{ccc} 1 & 2 & 1 \end{array}\right], \qquad \theta > 0,$$

where e_1 and e_3 are the pendant edges.

Proof. First suppose that $n \geq 5$ and that $\mathcal{T}$ is *not* the star. Then, necessarily, there are two pendant vertices u and v joined by a path of length $l \geq 3$. Suppose that u_0 is adjacent to u and v_0 is adjacent to v:



By Theorem 2.2, in $\mathcal{T}$ the set of vertices consisting of u_0 and its branches containing neither u nor v has cardinality $(n-2)/2$, and similarly for v_0. If $l \geq 4$, we get a contradiction to the number of vertices in $\mathcal{T}$. Hence, $\mathcal{T}$ can have paths of length at most three. It follows that in $\mathcal{T}$ there are exactly $(n-2)/2$ vertices at u_0, so n has to be even, and $\mathcal{T}$ has the following structure:



The matrix which must then be considered is, by Theorem 2.1,

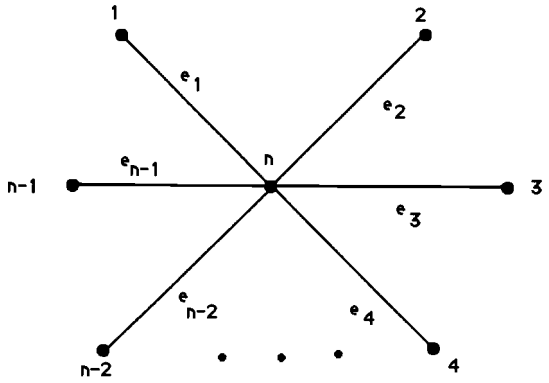
$$\left[\begin{array}{c|c} nI - J & -\frac{n^2}{4}\mathbf{1} \\ \hline -\mathbf{1^T} & \frac{n^2}{4} \end{array}\right].$$

The Schur complement for the $(n-1, n-1)$-entry is given by

$$\begin{aligned} \frac{n^2}{4}\left[1 - \mathbf{1}^T(nI-J)^{-1}\mathbf{1}\right] &= \frac{n^2}{4}\left[1 - \mathbf{1}^T\left(\frac{1}{n}I + \frac{1}{2n}J\right)\mathbf{1}\right] \\ &= \frac{n^2}{4}\left(1 - \frac{n-2}{2}\right) \\ &= \frac{(4-n)n^2}{8} \;<\; 0. \end{aligned}$$

Thus, this tree does not yield an M-matrix so that, if $n \geq 5$, only the star might admit an M-matrix weighting for $L^{\#}$.

We now find the admissible weightings for the star on n vertices:



We want weights $w(e_i)$, $1 \le i \le n-1$, such that

$$(nI - J) \begin{bmatrix} \frac{1}{w(e_1)} \\ \vdots \\ \frac{1}{w(e_n)} \end{bmatrix} = y$$

for some nonzero nonnegative vector y or, equivalently,

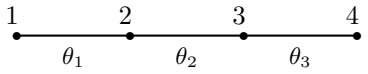
$$\begin{bmatrix} \frac{1}{w(e_1)} \\ \vdots \\ \frac{1}{w(e_n)} \end{bmatrix} = \frac{1}{n}(I + J)y$$

for some nonzero nonnegative vector y. Consequently, we have to select the weights so that

$$w(e_i) = \frac{1}{y_i + \sum_{j=1}^{n-1} y_j}, \qquad i = 1, \ldots, n-1,$$

for some vector y as mentioned.

Finally, to check the only tree which is *not* a star and has an admissible weighting, we consider the path on four vertices:



By Theorem 2.1, we have that

$$\begin{bmatrix} 3 & -4 & -1 \\ -1 & 4 & -1 \\ -1 & -4 & 3 \end{bmatrix} \begin{bmatrix} 1/\theta_1 \\ 1/\theta_2 \\ 1/\theta_3 \end{bmatrix} \ge 0.$$

But

$$A := \begin{bmatrix} 3 & -4 & -1 \\ -1 & 4 & -1 \\ -1 & -4 & 3 \end{bmatrix}$$

is a singular and irreducible M-matrix and therefore the only nonzero nonnegative (column) vectors which it maps to nonnegative vectors are nonnegative nonzero null vectors. As the nullspace of A is spanned by the vector $[2\ 1\ 2]^T$, we see that

$$\begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \end{bmatrix} = \theta \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$$

for some $\theta > 0$. □

REFERENCES

[1] A. Ben-Israel and T. N. Greville, *Generalized Inverses: Theory and Applications*, Academic Press, New York, 1973.

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

[3] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*, Dover, New York, 1991.

[4] D. E. Crabtree, *Applications of M–matrices to nonnegative matrices*, Duke Math. J., 33 (1996), pp. 197–208.

[5] E. Deutsch and M. Neumann, *Derivatives of the Perron root at an essentially nonnegative matrix and the group inverse of an M–matrix*, J. Math. Anal. Appl., 102 (1984), pp. 1–29.

[6] S. J. Kirkland, M. Neumann, and B. Shader, *Bounds on the subdominant eigenvalue involving group inverses with applications to graphs*, Czechoslovak Math. J., (1998), to appear.

[7] S. J. Kirkland, M. Neumann, and B. Shader, *Distances in weighted trees and group inverses of Laplacian matrices*, SIAM J. Matrix. Anal. Appl., 18 (1997), pp. 827–841.

[8] S. J. Kirkland, M. Neumann, and B. Shader, *On a bound on algebraic connectivity: The case of equality*, Czechoslovak Math. J., (1998), to appear.

[9] G. P. H. Styan and G. Subak-Sharpe, *Inequalities and equalities associated with the Campbell–Youla generalized inverse of the indefinite admittance matrix of resistive networks*, Linear Algebra Appl., 250 (1997), pp. 349–370.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 235–252, January 1998

016

PARAMETER ESTIMATION IN THE PRESENCE OF BOUNDED DATA UNCERTAINTIES*

S. CHANDRASEKARAN†, G. H. GOLUB‡, M. GU§, AND A. H. SAYED¶

Abstract. We formulate and solve a new parameter estimation problem in the presence of data uncertainties. The new method is suitable when a priori bounds on the uncertain data are available, and its solution leads to more meaningful results, especially when compared with other methods such as total least-squares and robust estimation. Its superior performance is due to the fact that the new method guarantees that the effect of the uncertainties will never be unnecessarily overestimated, beyond what is reasonably assumed by the a priori bounds. A geometric interpretation of the solution is provided, along with a closed form expression for it. We also consider the case in which only selected columns of the coefficient matrix are subject to perturbations.

Key words. least-squares estimation, regularized least-squares, ridge regression, total least-squares, robust estimation, modeling errors, secular equation

AMS subject classifications. 15A06, 65F05, 65F10, 65F35, 65K10, 93C41, 93E10, 93E24

PII. S0895479896301674

1. Introduction.

The central problem in estimation is to recover, to good accuracy, a set of unobservable parameters from corrupted data. Several optimization criteria have been used for estimation purposes over the years, but the most important, at least in the sense of having had the most applications, are criteria that are based on quadratic cost functions. The most striking among these is the linear least-squares criterion, which was first developed by Gauss (ca. 1795) in his work on celestial mechanics. Since then, it has enjoyed widespread popularity in many diverse areas as a result of its attractive computational and statistical properties (see, e.g., [4, 8, 10, 13]). Among these attractive properties, the most notable are the facts that least-squares solutions can be explicitly evaluated in closed forms, they can be recursively updated as more input data is made available, and they are also maximum likelihood estimators in the presence of normally distributed measurement noise.

Alternative optimization criteria, however, have been proposed over the years including, among others, regularized least-squares [4], ridge regression [4, 10], total least-squares [2, 3, 4, 7], and robust estimation [6, 9, 12, 14]. These different formulations allow, in one way or another, incorporation of further a priori information about the unknown parameter into the problem statement. They are also more effective in the presence of data errors and incomplete statistical information about the exogenous signals (or measurement errors).

Among the most notable variations is the total least-squares (TLS) method, also known as orthogonal regression or errors-in-variables method in statistics and system

*Received by the editors April 8, 1996; accepted for publication (in revised form) by N. J. Higham February 13, 1997.

http://www.siam.org/journals/simax/19-1/30167.html

†Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93016 (shiv@ ece.ucsb.edu).

‡Department of Computer Science, Stanford University, Stanford, CA 94305 (golub@ sccm.stanford.edu).

§Department of Mathematics, University of California, Los Angeles, CA 90095 (mgu@ math.ucla.edu).

¶Department of Electrical Engineering, University of California, Los Angeles, CA 90095 (sayed@ ee.ucla.edu). The work of this author was supported in part by National Science Foundation award MIP-9796147.

identification [11]. In contrast to the standard least-squares problem, the TLS formulation allows for errors in the data matrix. But it still exhibits certain drawbacks that degrade its performance in practical situations. In particular, it may unnecessarily overemphasize the effect of noise and uncertainties and can, therefore, lead to overly conservative results.

More specifically, assume $A \in \mathbf{R}^{m\times n}$ is a given full rank matrix with $m \geq n$, $b \in \mathbf{R}^m$ is a given vector, and consider the problem of solving the inconsistent linear system $A\hat{x} \approx b$ in the least-squares sense. The TLS solution assumes data uncertainties in A and proceeds to correct A and b by replacing them by their projections, $\hat{A}$ and $\hat{b}$, onto a specific subspace and by solving the consistent linear system of equations $\hat{A}\hat{x} = \hat{b}$. The spectral norm of the correction $(A - \hat{A})$ in the TLS solution is bounded by the smallest singular value of $[A\ b]$. While this norm might be small for vectors b that are close enough to the range space of A, it need not always be so. In other words, the TLS solution may lead to situations in which the correction term is unnecessarily large.

Consider, for example, a situation in which the uncertainties in A are very small, say, A is almost known exactly. Assume further that b is far from the column space of A. In this case, it is not difficult to visualize that the TLS solution will need to rotate (A, b) into $(\hat{A}, \hat{b})$ and may therefore end up with an overly corrected approximant for A, despite the fact that A is almost exact.

These facts motivate us to introduce a new parameter estimation formulation with prior bounds on the size of the allowable corrections to the data. More specifically, we formulate and solve a new estimation problem that is more suitable for scenarios in which a priori bounds on the uncertain data are known. The solution leads to more meaningful results in the sense that it guarantees that the effect of the uncertainties will never be unnecessarily overestimated, beyond what is reasonably assumed by the a priori bounds.

We note that, while preparing this paper, the related work [1] has come to our attention, where the authors have independently formulated and solved a similar estimation problem by using (convex) semidefinite programming techniques and interior-point methods. The resulting computational complexity of the proposed solution is $O(nm^2 + m^{3.5})$, where n is the smaller matrix dimension.

The solution proposed in this paper proceeds by first providing a geometric formulation of the problem, followed by an algebraic derivation that establishes that the optimal solution can in fact be obtained by solving a related regularized problem. The parameter of the regularization step is further shown to be obtained as the unique positive root of a secular equation and as a function of the given data. In this sense, the new formulation turns out to provide automatic regularization and, hence, has some useful regularization properties: the regularization parameter is not selected by the user but rather determined by the algorithm. Our solution involves an SVD step, and its computational complexity amounts to $O(mn^2 + n^3)$, where n is again the smaller matrix dimension. A summary of the problem and its solution is provided in section 3.4. (Other problem formulations are studied in [15].)

2. Problem formulation. Let $A \in \mathbf{R}^{m\times n}$ be a given matrix with $m \geq n$ and $b \in \mathbf{R}^m$ a given vector, both of which are assumed to be linearly related via an unknown vector of parameters $x \in \mathbf{R}^n$,

$$b = Ax + v\ . \tag{2.1}$$

The vector $v \in \mathbf{R}^m$ denotes measurement noise and it explains the mismatch between Ax and the given vector (or observation) b.

We assume that the "true" coefficient matrix is $A + \delta A$ and that we only know an upper bound on the 2-induced norm of the perturbation δA,

$$\|\delta A\|_2 \ \leq \ \eta \ , \tag{2.2}$$

with η being known. Likewise, we assume that the "true" observation vector is $b + \delta b$ and that we know an upper bound η_b on the Euclidean norm of the perturbation δb,

$$\|\delta b\|_2 \ \leq \ \eta_b \ . \tag{2.3}$$

We then pose the problem of finding an estimate that performs "well" for any allowed perturbation $(\delta A, \delta b)$. More specifically, we pose the following min-max problem.

PROBLEM 1. *Given $A \in \mathbf{R}^{m\times n}$, with $m \geq n$, $b \in \mathbf{R}^m$, and nonnegative real numbers (η, η_b), determine, if possible, an $\hat{x}$ that solves*

$$\min_{\hat{x}} \max\big\{ \| \, (A + \delta A)\, \hat{x} - (b + \delta b)\|_2 : \ \|\delta A\|_2 \leq \eta, \ \|\delta b\|_2 \leq \eta_b \big\} \ . \tag{2.4}$$

The situation is depicted in Fig. 2.1. Any particular choice for $\hat{x}$ would lead to many residual norms,

$$\| \, (A + \delta A)\, \hat{x} - (b + \delta b)\|_2 \ ,$$

one for each possible choice of A in the disc $(A + \delta A)$ and b in the disc $(b + \delta b)$. A second choice for $\hat{x}$ would lead to other residual norms, the maximum value of which need not be the same as the first choice. We want to choose an estimate $\hat{x}$ that minimizes the maximum possible residual norm. This is depicted in Fig. 2.2 for two choices, say $\hat{x}_1$ and $\hat{x}_2$. The curves show the values of the residual norms as a function of $(A + \delta A, b + \delta b)$.

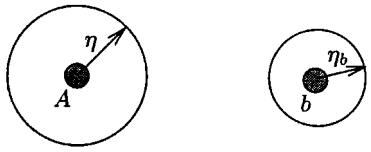


FIG. 2.1. *Geometric interpretation of the new least-squares formulation.*

We note that if $\eta = 0 = \eta_b$, then problem (2.4) reduces to a standard least-squares problem. Therefore we shall assume throughout that $\eta > 0$. (It will turn out that the solution to the above min-max problem is independent of η_b.)

2.1. A geometric interpretation. The min-max problem admits an interesting geometric formulation that highlights some of the issues involved in its solution.

For this purpose, and for the sake of illustration, assume we have a unit-norm vector b, $\|b\|_2 = 1$, with no uncertainties in it ($\eta_b = 0$). Assume further that A is simply a column vector, say a, with $\eta \neq 0$. That is, only A is assumed to be uncertain with perturbations that are bounded by η in magnitude (as in (2.2)). Now consider problem (2.4) in this context, which reads as follows:

$$\min_{\hat{x}} \left(\max_{\|\delta a\|_2 \leq \eta} \| \, (a + \delta a)\, \hat{x} - b\|_2 \right) \ . \tag{2.5}$$

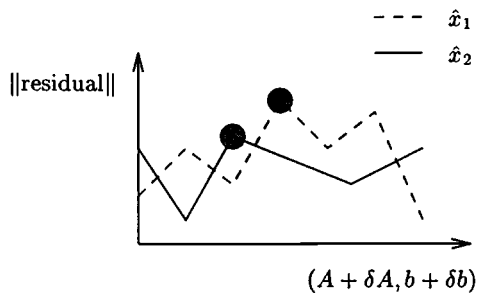


FIG. 2.2. *Two illustrative residual-norm curves.*

This situation is depicted in Fig. 2.3. The vectors a and b are indicated in thick black lines. The vector a is shown in the horizontal direction and a circle of radius η around its vertex indicates the set of all possible vertices for $a + \delta a$.

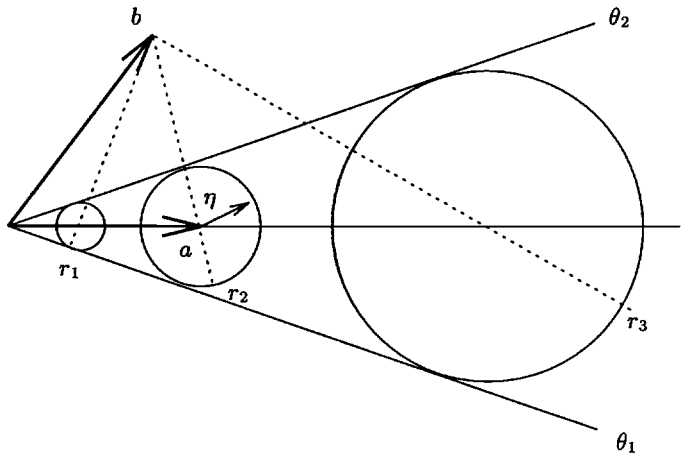


FIG. 2.3. *Geometric construction of the solution for a simple example.*

For any $\hat{x}$ that we pick, the set $\{(a+\delta a)\hat{x}\}$ describes a disc of center $a\hat{x}$ and radius $\eta\hat{x}$. This is indicated in the figure by the largest rightmost circle, which corresponds to a choice of a positive $\hat{x}$ that is larger than one. The vector in $\{(a + \delta a)\hat{x}\}$ that is furthest away from b is the one obtained by drawing a line from b through the center of the rightmost circle. The intersection of this line with the circle defines a residual vector r_3 whose norm is the largest among all possible residual vectors in the set $\{(a + \delta a)\hat{x}\}$.

Likewise, if we draw a line from b that passes through the vertex of a, it will intersect the circle at a point that defines a residual vector r_2. This residual will have the largest norm among all residuals that correspond to the particular choice $\hat{x} = 1$.

More generally, any $\hat{x}$ that we pick will determine a circle, and the corresponding largest residual is obtained by finding the furthest point on the circle from b. This is the point where the line that passes through b and the center of the circle intersects the circle on the other side of b.

We need to pick an $\hat{x}$ that minimizes the largest residual. For example, it is clear from the figure that the norm of r_3 is larger than the norm of r_2. The claim is that

in order to minimize the largest residual we need to proceed as follows: we drop a perpendicular from b to the lower tangent line denoted by θ_1. This perpendicular intersects the horizontal line in a point where we draw a new circle (the leftmost circle) that is tangent to both θ_1 and θ_2. This circle corresponds to a choice of $\hat{x}$ such that the furthest point on it from b is the foot of the perpendicular from b to θ_1. The residual indicated by r_1 corresponds to the desired solution (it has the minimum norm among the largest residuals).

To verify this claim, we refer to Fig. 2.4, where we have only indicated two circles, the circle that leads to a largest residual that is orthogonal to θ_1 and a second circle to its left. For this second leftmost circle, we denote its largest residual by r_4. We also denote the segment that connects b to the point of tangency of this circle with θ_1 by r. It is clear that r is larger than r_1 since r and r_1 are the sides of a right triangle. It is also clear that r_4 is larger than r by construction. Hence, r_4 is larger than r_1. A similar argument will show that r_1 is smaller than residuals that result from circles to its right.

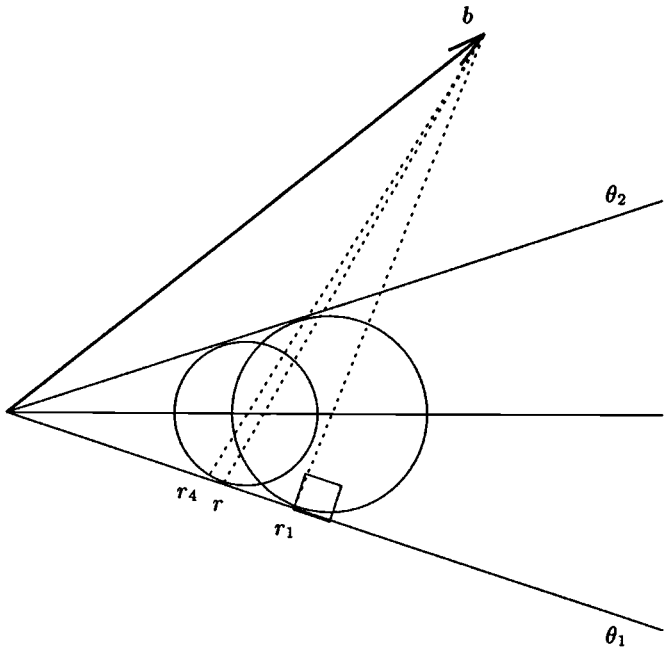


FIG. 2.4. *Geometric construction of the solution for a simple example.*

The above argument shows that the minimizing solution can be obtained as follows: drop a perpendicular from b to θ_1. Pick the point where the perpendicular meets the horizontal line and draw a circle that is tangent to both θ_1 and θ_2. Its radius will be $\eta\hat{x}$, where $\hat{x}$ is the optimal solution. Also, the foot of the perpendicular on θ_1 will be the optimal $\hat{b}$.

The projection $\hat{b}$ (and consequently the solution $\hat{x}$) will be nonzero as long as b is not orthogonal to the direction θ_1. This imposes a condition on η. Indeed, the direction θ_1 will be orthogonal to b only when η is large enough. This requires that the circle centered around a has radius $a^T b$, which is the length of the projection of a onto the unit norm vector b. This is depicted in Fig. 2.5.

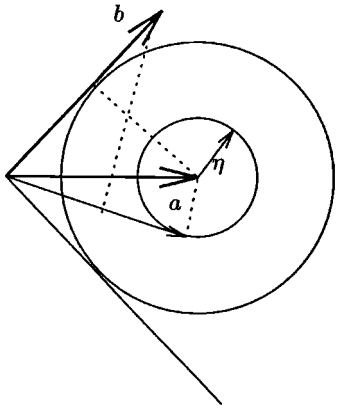


FIG. 2.5. *Geometric condition for a nonzero solution.*

Hence, the largest value that can be allowed for η in order to have a nonzero solution $\hat{x}$ is

$$\eta \;<\; |a^T b| \;.$$

Indeed, if η were larger than or equal to this value, then the vector in the set $(a + \delta a)$ that would always lead to the maximum residual norm is the one that is orthogonal to b, in which case the solution will be zero again. The same geometric argument will lead to a similar conclusion had we allowed for uncertainties in b as well.

For a nonunity b, the upper bound on η would take the form

$$\eta \;<\; \frac{|a^T b|}{\|b\|_2} \;.$$

We shall see that in the general case a similar bound holds, for nonzero solutions, and is given by

$$\eta \;<\; \frac{\|A^T b\|_2}{\|b\|_2} \;.$$

We now proceed to an algebraic solution of the min-max problem. A final statement of the form of the solution is given in section 3.4.

3. Reducing the min-max problem to a minimization problem. We start by showing how to reduce the min-max problem (2.4) to a standard minimization problem. To begin with, we note that

$$\begin{aligned} \|\,(A+\delta A)\,\hat{x} - (b+\delta b)\|_2 \;&\leq\; \|A\hat{x} - b\|_2 + \|\delta A\|_2 \cdot \|\hat{x}\|_2 \;+\; \|\delta b\|_2 \;, \\ &\leq\; \|A\hat{x} - b\|_2 + \eta\|\hat{x}\|_2 \;+\; \eta_b \;, \end{aligned}$$

which provides an upper bound for $\|\,(A+\delta A)\,\hat{x} - (b+\delta b)\|_2$. But this upper bound is in fact achievable, i.e., there exist $(\delta A, \delta b)$ for which

$$\|\,(A+\delta A)\,\hat{x} - (b+\delta b)\|_2 = \|A\hat{x} - b\|_2 + \eta\|\hat{x}\|_2 + \eta_b \;.$$

To see that this is indeed the case, choose δA as the rank one matrix

$$\delta A^o = \frac{(A\hat{x} - b)}{\|A\hat{x} - b\|_2} \frac{\hat{x}^T}{\|\hat{x}\|_2}\eta \ ,$$

and choose δb as the vector

$$\delta b^o = -\frac{(A\hat{x} - b)}{\|A\hat{x} - b\|_2} \eta_b \ .$$

For these choices of perturbations in A and b, it follows that

$$(A\hat{x} - b) \ , \quad \delta A^o \hat{x} \ , \quad \text{and} \quad \delta b^o \ ,$$

are collinear vectors that point in the same direction. Hence,

$$\begin{aligned} \| \left(A + \delta A^o\right) \hat{x} - (b + \delta b^o)\|_2 &= \|(A\hat{x} - b) + \delta A^o \hat{x} - \delta b^o\|_2 \ , \\ &= \|A\hat{x} - b\|_2 + \|\delta A^o \hat{x}\|_2 + \|\delta b^o\|_2 \ , \\ &= \|A\hat{x} - b\|_2 + \eta\|\hat{x}\|_2 + \eta_b \ , \end{aligned}$$

which is the desired upper bound. We therefore conclude that

$$\max_{\|\delta A\|_2 \leq \eta \ , \|\delta b\|_2 \leq \eta_b} \| \left(A + \delta A\right) \hat{x} - (b + \delta b)\|_2 = \|A\hat{x} - b\|_2 + \eta\|\hat{x}\|_2 + \eta_b \ , \tag{3.1}$$

which establishes the following result.

LEMMA 3.1. *The min-max problem* (2.4) *is equivalent to the following minimization problem. Given $A \in \mathbf{R}^{m\times n}$, with $m \geq n$, $b \in \mathbf{R}^m$, and nonnegative real numbers (η, η_b), determine, if possible, an $\hat{x}$ that solves*

$$\min_{\hat{x}} \ (\|A\hat{x} - b\|_2 + \eta\|\hat{x}\|_2 + \eta_b) \ . \tag{3.2}$$

3.1. Solving the minimization problem. To solve (3.2), we define the cost function

$$\mathcal{L}(\hat{x}) = \|A\hat{x} - b\|_2 + \eta\|\hat{x}\|_2 + \eta_b \ .$$

It is easy to check that $\mathcal{L}(\hat{x})$ is a convex continuous function in $\hat{x}$, and hence, any local minimum of $\mathcal{L}(\hat{x})$ is also a global minimum. But at any local minimum of $\mathcal{L}(\hat{x})$, it either holds that $\mathcal{L}(\hat{x})$ is not differentiable or its gradient $\bigtriangledown\mathcal{L}(\hat{x})$ is 0. In particular, note that $\mathcal{L}(\hat{x})$ is not differentiable only at $\hat{x} = 0$ and at any $\hat{x}$ that satisfies $A\hat{x} - b = 0$.

We first consider the case in which $\mathcal{L}(\hat{x})$ is differentiable and, hence, the gradient of $\mathcal{L}(\hat{x})$ exists and is given by

$$\begin{aligned} \bigtriangledown\mathcal{L}(\hat{x}) &= \frac{1}{\|A\hat{x} - b\|_2} A^T \left(A\hat{x} - b\right) + \frac{\eta}{\|\hat{x}\|_2}\hat{x} \ , \\ &= \frac{1}{\|A\hat{x} - b\|_2} \left(\left(A^T A + \alpha I\right)\hat{x} - A^T b\right) \ , \end{aligned}$$

where we have introduced the positive real number

$$\alpha = \frac{\eta\|A\hat{x} - b\|_2}{\|\hat{x}\|_2} \ . \tag{3.3}$$

By setting $\bigtriangledown \mathcal{L}(\hat{x}) = 0$ we obtain that any stationary solution $\hat{x}$ of $\mathcal{L}(\hat{x})$ is given by

$$\hat{x} = \left(A^T A + \alpha I\right)^{-1} A^T b \, . \tag{3.4}$$

We still need to determine the parameter α that corresponds to $\hat{x}$ and which is defined in (3.3).

To solve for α, we introduce the singular value decomposition (SVD) of A,

$$A = U \left[\begin{array}{c} \Sigma \\ 0 \end{array} \right] V^T \, , \tag{3.5}$$

where $U \in \mathbf{R}^{m \times m}$ and $V \in \mathbf{R}^{n \times n}$ are orthogonal, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is diagonal, with

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$$

being the singular values of A. We further partition the vector $U^T b$ into

$$\left[\begin{array}{c} b_1 \\ b_2 \end{array} \right] = U^T b, \tag{3.6}$$

where $b_1 \in \mathbf{R}^n$ and $b_2 \in \mathbf{R}^{m-n}$.

In this case, the expression (3.4) for $\hat{x}$ can be rewritten in the equivalent form

$$\hat{x} = V(\Sigma^2 + \alpha I)^{-1} \Sigma b_1 \, , \tag{3.7}$$

and, hence,

$$\|\hat{x}\|_2 = \|\Sigma \left(\Sigma^2 + \alpha I\right)^{-1} b_1\|_2 \, .$$

Likewise,

$$\begin{aligned} b - A\hat{x} &= U \left(U^T b - \left(\begin{array}{c} \Sigma \\ 0 \end{array} \right) \left(\Sigma^2 + \alpha I\right)^{-1} \Sigma b_1 \right) \, , \\ &= U \left[\begin{array}{c} b_1 - \Sigma^2 \left(\Sigma^2 + \alpha I\right)^{-1} b_1 \\ b_2 \end{array} \right] \, , \\ &= U \left[\begin{array}{c} \alpha \left(\Sigma^2 + \alpha I\right)^{-1} b_1 \\ b_2 \end{array} \right] \, , \end{aligned}$$

which shows that

$$\|b - A\hat{x}\|_2 = \sqrt{\|b_2\|_2^2 + \alpha^2 \| \left(\Sigma^2 + \alpha I\right)^{-1} b_1\|_2^2} \, .$$

Therefore, (3.3) for α reduces to the following nonlinear equation that is only a function of α and the given data (A, b, η):

$$\alpha = \frac{\eta \sqrt{\|b_2\|_2^2 + \alpha^2 \| \left(\Sigma^2 + \alpha I\right)^{-1} b_1\|_2^2}}{\|\Sigma \left(\Sigma^2 + \alpha I\right)^{-1} b_1\|_2} \, . \tag{3.8}$$

Note that only the norm of b_2, and not b_2 itself, is needed in the above expression.

Remark. We have assumed in the derivation so far that A is full rank. If this were not the case, i.e., if A (and hence Σ) were singular, then (3.8) can be reduced to an equation of the same form but with a nonsingular Σ of smaller dimension. Indeed, if we partition

$$\Sigma = \begin{bmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{bmatrix},$$

where $\hat{\Sigma} \in \mathbf{R}^{k\times k}$ is nonsingular, and let $\hat{b}_1 \in \mathbf{R}^k$ be the first k components of b_1, $\tilde{b}_1 \in \mathbf{R}^{n-k}$ be the last $n-k$ components of b_1, and let

$$\|\hat{b}_2\|_2^2 = \|b_2\|_2^2 + \|\tilde{b}_1\|_2^2,$$

then (3.8) reduces to

$$\alpha = \frac{\eta\sqrt{\|\hat{b}_2\|_2^2 + \alpha^2\|\left(\hat{\Sigma}^2 + \alpha I\right)^{-1}\hat{b}_1\|_2^2}}{\|\hat{\Sigma}\left(\hat{\Sigma}^2 + \alpha I\right)^{-1}\hat{b}_1\|_2}, \tag{3.9}$$

which is the same form as (3.8). From now on, we assume that A is full rank and, hence, Σ is invertible:

A full rank is a standing assumption in what follows .

3.2. The secular equation. Define the nonlinear function in α,

$$\mathcal{G}(\alpha) = b_1^T\left(\Sigma^2 - \eta^2 I\right)\left(\Sigma^2 + \alpha I\right)^{-2} b_1 - \frac{\eta^2}{\alpha^2}\|b_2\|_2^2 . \tag{3.10}$$

It is clear that α is a positive solution to (3.8) if, and only if, it is a positive root of $\mathcal{G}(\alpha)$. Following [4], we refer to the equation

$$\mathcal{G}(\alpha) = 0 \tag{3.11}$$

as a *secular* equation.

The function $\mathcal{G}(\alpha)$ has several useful properties that will allow us to provide conditions for the existence of a unique positive root α. We start with the following result.

LEMMA 3.2. *The function $\mathcal{G}(\alpha)$ in* (3.10) *can have at most one positive root. In addition, if $\hat{\alpha} > 0$ is a root of $\mathcal{G}(\alpha)$, then $\hat{\alpha}$ is a simple root and $\mathcal{G}'(\hat{\alpha}) > 0$.*

Proof. We prove the second conclusion first. Partition

$$\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \in \mathbf{R}^{(n+1)\times(n+1)},$$

where the diagonal entries of $\Sigma_1 \in \mathbf{R}^{k\times k}$ are those of Σ that are larger than η, and the diagonal entries of $\Sigma_2 \in \mathbf{R}^{(n+1-k)\times(n+1-k)}$ are the remaining diagonal entries of Σ and one 0. It follows that (in terms of the 2-induced norm for the diagonal matrices $(\Sigma_2^2 + \alpha I)$ and $(\Sigma_1^2 + \alpha I)$)

$$\|\Sigma_2^2 + \alpha I\|_2 \cdot \|\left(\Sigma_1^2 + \alpha I\right)^{-1}\|_2 < 1 \tag{3.12}$$

for all $\alpha > 0$.

Let $u \in \mathbf{R}^k$ be the first k components of $\sqrt{\Sigma^2 - \eta^2 I} \cdot b_1$ and let $v \in \mathbf{R}^{n+1-k}$ be the last $n+1-k$ components of

$$\begin{bmatrix} \sqrt{\eta^2 I - \Sigma^2} & 0 \\ 0 & \eta \end{bmatrix} \begin{bmatrix} b_1 \\ \|b_2\|_2 \end{bmatrix} .$$

It follows that we can rewrite $\mathcal{G}(\alpha)$ as the difference

$$\mathcal{G}(\alpha) = u^T \left(\Sigma_1^2 + \alpha I\right)^{-2} u - v^T \left(\Sigma_2^2 + \alpha I\right)^{-2} v$$

and, consequently,

$$\mathcal{G}'(\alpha) = -2 \left(u^T \left(\Sigma_1^2 + \alpha I\right)^{-3} u - v^T \left(\Sigma_2^2 + \alpha I\right)^{-3} v \right) .$$

Let $\hat{\alpha} > 0$ be a root of $\mathcal{G}(\alpha)$. This means that

$$u^T \left(\Sigma_1^2 + \hat{\alpha} I\right)^{-2} u = v^T \left(\Sigma_2^2 + \hat{\alpha} I\right)^{-2} v ,$$

which leads to the following sequence of inequalities:

$$\begin{aligned} u^T \left(\Sigma_1^2 + \hat{\alpha} I\right)^{-3} u &\leq \| \left(\Sigma_1^2 + \hat{\alpha} I\right)^{-1} \|_2 \cdot u^T \cdot \left(\Sigma_1^2 + \hat{\alpha} I\right)^{-2} u \\ &= \| \left(\Sigma_1^2 + \hat{\alpha} I\right)^{-1} \|_2 \cdot v^T \cdot \left(\Sigma_2^2 + \hat{\alpha} I\right)^{-2} v \\ &< \frac{1}{\| \left(\Sigma_2^2 + \alpha I\right) \|_2} \cdot v^T \cdot \left(\Sigma_2^2 + \hat{\alpha} I\right)^{-2} v \\ &\leq v^T \left(\Sigma_2^2 + \hat{\alpha} I\right)^{-3} v . \end{aligned}$$

Combining this relation with the expression for $\mathcal{G}'(\alpha)$, it immediately follows that $\mathcal{G}'(\hat{\alpha}) > 0$. Consequently, $\hat{\alpha}$ must be a simple root of $\mathcal{G}(\alpha)$.

Furthermore, we note that $\mathcal{G}(\alpha)$ is a sum of $n+1$ rational functions in α and hence can have only a finite number of positive roots. In the following we show by contradiction that $\mathcal{G}(\alpha)$ can have no positive roots other than $\hat{\alpha}$. Assume to the contrary that $\hat{\alpha}_1$ is another positive root of $\mathcal{G}(\alpha)$. Without loss of generality, we further assume that $\hat{\alpha} < \hat{\alpha}_1$ and that $\mathcal{G}(\alpha)$ does not have any root within the open interval $(\hat{\alpha}, \hat{\alpha}_1)$. It follows from the above proof that

$$\mathcal{G}'(\hat{\alpha}) > 0 \quad \text{and} \quad \mathcal{G}'(\hat{\alpha}_1) > 0 .$$

But this implies that $\mathcal{G}(\alpha) > 0$ for α slightly larger than $\hat{\alpha}$ and $\mathcal{G}(\alpha) < 0$ for α slightly smaller than $\hat{\alpha}_1$, and consequently, $\mathcal{G}(\alpha)$ must have a root in the interval $(\hat{\alpha}, \hat{\alpha}_1)$; a contradiction to our assumptions. So $\mathcal{G}(\alpha)$ can have at most one positive root. ∎

Now we provide conditions for $\mathcal{G}(\alpha)$ to have a positive root. (The next result was in fact suggested earlier by the geometric argument of Fig. 2.3.) Note that $A\hat{x}$ can be written as

$$A\hat{x} = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T \hat{x} .$$

Therefore solving $A\hat{x} = b$, when possible, is equivalent to solving

$$\begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T \hat{x} \;=\; U^T b \;=\; \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} .$$

This shows that a necessary and sufficient condition for b to belong to the column span of A is $b_2 = 0$.

LEMMA 3.3. *Assume $\eta > 0$ (a standing assumption) and $b_2 \neq 0$, i.e., b does not belong to the column span of A. Then the function $\mathcal{G}(\alpha)$ in* (3.10) *has a unique positive root if and only if*

$$\eta < \frac{\|A^T b\|_2}{\|b\|_2} \,. \tag{3.13}$$

Proof. We note that

$$\lim_{\alpha \to 0^+} \left(\alpha^2 \mathcal{G}(\alpha)\right) = -\eta^2 \|b_2\|_2^2 < 0 \,,$$

and that

$$\begin{aligned} \lim_{\alpha \to +\infty} \left(\alpha^2 \mathcal{G}(\alpha)\right) &= b_1^T \left(\Sigma^2 - \eta^2 I\right) b_1 - \eta^2 \|b_2\|_2^2 \,, \\ &= \|A^T b\|_2^2 - \eta^2 \|b\|_2^2 \,, \\ &= b_1^T \Sigma^2 b_1 - \eta^2 \|b\|_2^2 \,. \end{aligned} \tag{3.14}$$

First we assume that condition (3.13) holds. It follows then that $\mathcal{G}(\alpha)$ changes sign on the interval $(0, +\infty)$ and therefore has to have a positive root. By Lemma 3.2 this positive root must also be unique.

On the other hand, assume that

$$\eta > \frac{\|A^T b\|_2}{\|b\|_2} \,.$$

This condition implies, in view of (3.14), that $\mathcal{G}(\alpha) < 0$ for sufficiently large α. We now show by contradiction that $\mathcal{G}(\alpha)$ does not have a positive root. Assume to the contrary that $\hat{\alpha}$ is a positive root of $\mathcal{G}(\alpha)$. It then follows from Lemma 3.2 that $\mathcal{G}(\alpha)$ is positive for α slightly larger than $\hat{\alpha}$ since $\mathcal{G}'(\hat{\alpha}) > 0$, and hence $\mathcal{G}(\alpha)$ must have a root in $(\hat{\alpha}, +\infty)$, which is a contradiction according to Lemma 3.2. Hence $\mathcal{G}(\alpha)$ does not have a positive root in this case.

Finally, we consider the case

$$\eta = \frac{\|A^T b\|_2}{\|b\|_2} \,.$$

We also show by contradiction that $\mathcal{G}(\alpha)$ does not have a positive root. Assume to the contrary that $\hat{\alpha}$ is a positive root of $\mathcal{G}(\alpha)$. It then follows from Lemma 3.2 that $\hat{\alpha}$ must be a simple root and a continuous function of the coefficients in $\mathcal{G}(\alpha)$. In particular, $\hat{\alpha}$ is a continuous function of η. Now we slightly increase the value of η so that

$$\eta > \frac{\|A^T b\|_2}{\|b\|_2} \,.$$

By continuity, $\mathcal{G}(\alpha)$ has a positive root for such values of η, but we have just shown that for $\eta > \|A^T b\|_2 / \|b\|_2$ this is not possible. Hence, $\mathcal{G}(\alpha)$ does not have a positive root in this case either. ∎

We now consider the case $b_2 = 0$, i.e., b lies in the column span of A. This case arises, for example, when A is a square invertible matrix ($m = n$).

Define

$$\tau_1 = \frac{\|\Sigma^{-1} b_1\|_2}{\|\Sigma^{-2} b_1\|_2} \quad \text{and} \quad \tau_2 = \frac{\|\Sigma b_1\|_2}{\|b_1\|_2} \ .$$

It follows from $b_2 = 0$ that (cf. (3.13))

$$\tau_2 = \frac{\|A^T b\|_2}{\|b\|_2} \ .$$

Now note that

$$b_1^T b_1 = b_1^T \Sigma \Sigma^{-1} b_1 \ .$$

Therefore, by using the Cauchy–Schwarz inequality, we have

$$\|b_1\|_2 \|b_1\|_2 \ \leq \ \|\Sigma b_1\|_2 \ \|\Sigma^{-1} b_1\|_2 \ ,$$

and we obtain, after applying the Cauchy–Schwarz inequality one more time, that

$$\tau_2 = \frac{\|\Sigma b_1\|_2}{\|b_1\|_2} \geq \frac{\|b_1\|_2}{\|\Sigma^{-1} b_1\|_2} \geq \frac{\|\Sigma^{-1} b_1\|_2}{\|\Sigma^{-2} b_1\|_2} = \tau_1 \ . \tag{3.15}$$

LEMMA 3.4. *Assume $\eta > 0$ (a standing assumption) and $b_2 = 0$, i.e., b lies in the column span of A. Then the function $\mathcal{G}(\alpha)$ in* (3.10) *has a positive root if and only if*

$$\tau_1 < \eta < \tau_2 \ . \tag{3.16}$$

Proof. It is easy to check that

$$\lim_{\alpha \to 0^+} \mathcal{G}(\alpha) = \left(\tau_1^2 - \eta^2\right) b_1^T \Sigma^{-4} b_1 \ ,$$

and that

$$\lim_{\alpha \to +\infty} \left(\alpha^2 \mathcal{G}(\alpha)\right) = \left(\tau_2^2 - \eta^2\right) b_1^T b_1 \ .$$

If $\eta > \tau_2$, then

$$\lim_{\alpha \to 0^+} \mathcal{G}(\alpha) < 0 \quad \text{and} \quad \lim_{\alpha \to +\infty} \left(\alpha^2 \mathcal{G}(\alpha)\right) < 0 \ .$$

Arguments similar to those in the proof of Lemma 3.3 show that $\mathcal{G}(\alpha)$ does not have a positive root. Similarly $\mathcal{G}(\alpha)$ does not have a positive root if $\eta < \tau_1$. Continuity arguments similar to those in the proof of Lemma 3.3 show that $\mathcal{G}(\alpha)$ does not have a positive root if $\eta = \tau_2$ or τ_1.

However, if $\tau_1 < \eta < \tau_2$, then

$$\lim_{\alpha \to 0^+} \mathcal{G}(\alpha) < 0 \quad \text{and} \quad \lim_{\alpha \to +\infty} \left(\alpha^2 \mathcal{G}(\alpha)\right) > 0 \ .$$

So $\mathcal{G}(\alpha)$ must have a positive root. By Lemma 3.2 this positive root is unique. ∎

3.3. Finding the global minimum. We now show that whenever $\mathcal{G}(\alpha)$ has a positive root $\hat{\alpha}$, the corresponding vector $\hat{x}$ in (3.4) must be the global minimizer of $\mathcal{L}(\hat{x})$.

LEMMA 3.5. *Let $\hat{\alpha}$ be a positive root of $\mathcal{G}(\alpha)$ and let $\hat{x}$ be defined by* (3.4) *for $\alpha = \hat{\alpha}$. Then $\hat{x}$ is the global minimum of $\mathcal{L}(\hat{x})$.*

Proof. We first show that

$$\triangle\mathcal{L}(\hat{x}) > 0 \ ,$$

where $\triangle\mathcal{L}(\hat{x})$ is the Hessian of $\mathcal{L}$ at $\hat{x}$. We take the gradient of $\mathcal{L}$,

$$\bigtriangledown\mathcal{L}(\hat{x}) = \frac{1}{\|A\hat{x}-b\|_2} A^T \left(A\hat{x}-b\right) + \frac{\eta}{\|\hat{x}\|_2}\hat{x} \ .$$

Consequently,

$$\begin{aligned}\triangle\mathcal{L}(\hat{x}) = &\frac{1}{\|A\hat{x}-b\|_2} A^TA - \frac{1}{\|A\hat{x}-b\|_2^3}\left(A^TA\hat{x}-A^Tb\right)\left(A^TA\hat{x}-A^Tb\right)^T \\ &+\frac{\eta}{\|\hat{x}\|_2} I - \frac{\eta}{\|\hat{x}\|_2^3}\hat{x}\hat{x}^T \ .\end{aligned}$$

We now simplify this expression. It follows from (3.4) that

$$\left(A^TA+\hat{\alpha}I\right)\hat{x} = A^Tb \ ,$$

and, hence,

$$A^TA\hat{x} - A^Tb = -\hat{\alpha}\hat{x} \ .$$

Substituting this relation into the expression for the Hessian matrix $\triangle\mathcal{L}(\hat{x})$, and simplifying the resulting expression using (3.3), we obtain

$$\triangle\mathcal{L}(\hat{x}) = \frac{1}{\|A\hat{x}-b\|_2}\left(\left(A^TA+\hat{\alpha}I\right) - \frac{\hat{x}\hat{x}^T}{\hat{x}^T\hat{x}}\left(\hat{\alpha}+\eta^2\right)\right) \ .$$

Observe that the matrix $\left(A^TA+\hat{\alpha}I\right)$ is positive definite since $\hat{\alpha} > 0$. Hence $\triangle\mathcal{L}(\hat{x})$ can have at most one nonpositive eigenvalue. This implies that $\triangle\mathcal{L}(\hat{x})$ is positive definite if and only if $\det\left(\triangle\mathcal{L}(\hat{x})\right) > 0$. Indeed,

$$\begin{aligned}\frac{\det\left(\triangle\mathcal{L}(\hat{x})\right)\|A\hat{x}-b\|_2^n}{\det\left(A^TA+\hat{\alpha}I\right)} &= \det\left(I - \frac{\left(A^TA+\hat{\alpha}I\right)^{-1}\hat{x}\hat{x}^T}{\hat{x}^T\hat{x}}\left(\hat{\alpha}+\eta^2\right)\right) \\ &= 1 - \frac{\hat{x}^T\left(A^TA+\hat{\alpha}I\right)^{-1}\hat{x}}{\hat{x}^T\hat{x}}\left(\hat{\alpha}+\eta^2\right) \\ &= \frac{1}{\hat{x}^T\hat{x}}\left(\hat{x}^T\hat{x} - \left(\hat{\alpha}+\eta^2\right)\left(\hat{x}^T\left(A^TA+\hat{\alpha}I\right)^{-1}\hat{x}\right)\right) \ .\end{aligned}$$

The last expression can be further rewritten, using the SVD of A and (3.8),

$$\begin{aligned}\frac{\det\left(\triangle\mathcal{L}(\hat{x})\right)\|A\hat{x}-b\|_2^n}{\det\left(A^TA+\hat{\alpha}I\right)} = &\frac{1}{\hat{x}^T\hat{x}} b_1^T\Sigma^2\left(\Sigma^2+\hat{\alpha}I\right)^{-2} b_1 \\ &-\frac{\hat{\alpha}+\eta^2}{\hat{x}^T\hat{x}} b_1^T\Sigma^2\left(\Sigma^2+\hat{\alpha}I\right)^{-3} b_1\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\hat{x}^T\hat{x}} \frac{\eta^2 \left(\|b_2\|_2^2 + \hat{\alpha}^2 \| \left(\Sigma^2 + \hat{\alpha} I\right)^{-1} b_1 \|_2^2 \right)}{\hat{\alpha}^2} \\
&\quad - \frac{\hat{\alpha} + \eta^2}{\hat{x}^T\hat{x}} b_1^T \Sigma^2 \left(\Sigma^2 + \hat{\alpha} I\right)^{-3} b_1 \\
&= \frac{\hat{\alpha}}{\hat{x}^T\hat{x}} \left(\frac{\eta^2 \|b_2\|_2^2}{\hat{\alpha}^3} + b_1^T \left(\eta^2 - \Sigma^2\right) \left(\Sigma^2 + \hat{\alpha} I\right)^{-3} b_1 \right) .
\end{aligned}
$$

Comparing the last expression with the function $\mathcal{G}(\alpha)$ in (3.10), we finally have

$$
\frac{\det\left(\triangle \mathcal{L}(\hat{x})\right) \|A\hat{x} - b\|_2^n}{\det\left(A^T A + \hat{\alpha} I\right)} = \frac{\hat{\alpha}}{2\hat{x}^T\hat{x}} \mathcal{G}'(\hat{\alpha}) .
$$

By Lemma 3.2, we have that $\mathcal{G}'(\hat{\alpha}) > 0$. Consequently, $\triangle \mathcal{L}(\hat{x})$ must be positive definite, and hence $\hat{x}$ must be a local minimizer of $\mathcal{L}(\hat{x})$. Since $\mathcal{L}(\hat{x})$ is a convex function, this also means that $\hat{x}$ is a global minimizer of $\mathcal{L}(\hat{x})$. □

We still need to consider the points at which $\mathcal{L}(\hat{x})$ is not differentiable. These include $\hat{x} = 0$ and any solution of $A\hat{x} = b$.

Consider first the case $b_2 \neq 0$. This means that b does not belong to the column span of A and, hence, we only need to check $\hat{x} = 0$. If condition (3.13) holds, then it follows from Lemma 3.3 that $\mathcal{G}(\alpha)$ has a unique positive root $\hat{\alpha}$, and it follows from Lemma 3.5 that

$$
\hat{x} = \left(A^T A + \hat{\alpha} I\right)^{-1} A^T b
$$

is the global minimum. On the other hand, if condition (3.13) does not hold, then it follows from Lemma 3.3 that $\mathcal{G}(\alpha)$ does not have any positive root and hence

$$
\hat{x} = 0
$$

is the global minimum.

Now consider the case $b_2 = 0$, which means that b lies in the column span of A. In this case $\mathcal{L}(\hat{x})$ is not differentiable at both $\hat{x} = 0$ and $\hat{x} = V\Sigma^{-1}b_1 = A^\dagger b$. If condition (3.16) holds, then it follows from Lemma 3.4 that $\mathcal{G}(\alpha)$ has a unique positive root $\hat{\alpha}$ and it follows from Lemma 3.5 that

$$
\hat{x} = \left(A^T A + \hat{\alpha} I\right)^{-1} A^T b
$$

is the global minimum. On the other hand, if $\eta \leq \tau_1$, then

$$
\begin{aligned}
\mathcal{L}\left(V\Sigma^{-1}b_1\right) - \mathcal{L}(0) &= \eta \|\Sigma^{-1}b_1\|_2 - \|b_1\|_2 , \\
&\leq \|\Sigma^{-1}b_1\|_2 \left(\frac{\|\Sigma^{-1}b_1\|_2}{\|\Sigma^{-2}b_1\|_2} - \frac{\|b_1\|_2}{\|\Sigma^{-1}b_1\|_2} \right) , \\
&\leq 0 ,
\end{aligned}
$$

where we have used the Cauchy–Schwarz inequality. It follows that

$$
\hat{x} = V\Sigma^{-1}b_1
$$

is the global minimum in this case. Similarly, if $\eta \geq \tau_2$, then

$$
\hat{x} = 0
$$

is the global minimum.

We finally consider the degenerate case $\tau_1 = \tau_2 = \eta$. Under this condition, it follows from (3.15) that

$$\|\Sigma^{-1}b_1\|_2\|\Sigma b_1\|_2 = \|b_1\|_2 \cdot \|b_1\|_2 \ .$$

Hence,

$$\begin{aligned}\mathcal{L}\left(V\Sigma^{-1}b_1\right) - \mathcal{L}(0) &= \eta\|\Sigma^{-1}b_1\|_2 - \|b_1\|_2 \\ &= \frac{\|\Sigma^{-1}b_1\|_2}{\|b_1\|_2} \cdot \|\Sigma^{-1}b_1\|_2 - \|b_1\|_2 = 0 \ .\end{aligned}$$

This shows that $\mathcal{L}\left(V\Sigma^{-1}b_1\right) = \mathcal{L}(0)$. But since $\mathcal{L}(\hat{x})$ is a convex function in $\hat{x}$, we conclude that for any $\hat{x}$ that is a convex linear combination of 0 and $V\Sigma^{-1}b_1$, say

$$\hat{x} = \beta V\Sigma^{-1}b_1, \quad \text{for any } \ 0 \le \beta \ \le 1 \ , \tag{3.17}$$

we also obtain $\mathcal{L}(\hat{x}) = 0$. Therefore, when $\tau_1 = \tau_2 = \eta$ there are many solutions $\hat{x}$ and these are all scaled multiples of $V\Sigma^{-1}b_1$ as in (3.17).

3.4. Statement of the solution of the min-max problem. We collect, in the form of a theorem, the conclusions of our earlier analysis.

THEOREM 3.6. *Given $A \in \mathbf{R}^{m\times n}$, with $m \ge n$ and A full rank, $b \in \mathbf{R}^m$, and nonnegative real numbers (η, η_b). The following optimization problem*

$$\min_{\hat{x}} \max\left\{\|\left(A + \delta A\right)\hat{x} - \left(b + \delta b\right)\|_2 : \ \|\delta A\|_2 \le \eta, \ \|\delta b\|_2 \le \eta_b\right\} \tag{3.18}$$

always has a solution $\hat{x}$. The solution(s) can be constructed as follows.

- *Introduce the SVD of A,*

$$A = U \begin{bmatrix} \Sigma \\ 0 \end{bmatrix} V^T \ , \tag{3.19}$$

where $U \in \mathbf{R}^{m\times m}$ and $V \in \mathbf{R}^{n\times n}$ are orthogonal, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ is diagonal, with

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n > 0$$

being the singular values of A.

- *Partition the vector U^Tb into*

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = U^Tb, \tag{3.20}$$

where $b_1 \in \mathbf{R}^n$ and $b_2 \in \mathbf{R}^{m-n}$.

- *Introduce the secular function*

$$\mathcal{G}(\alpha) = b_1^T\left(\Sigma^2 - \eta^2 I\right)\left(\Sigma^2 + \alpha I\right)^{-2} b_1 - \frac{\eta^2}{\alpha^2}\|b_2\|_2^2 \ . \tag{3.21}$$

- *Define*

$$\tau_1 = \frac{\|\Sigma^{-1}b_1\|_2}{\|\Sigma^{-2}b_1\|_2} \quad \textit{and} \quad \tau_2 = \frac{\|A^Tb\|_2}{\|b\|_2} \ .$$

First case: b does not belong to the column span of A.

1. *If $\eta \geq \tau_2$, then the unique solution is $\hat{x} = 0$.*
2. *If $\eta < \tau_2$, then the unique solution is $\hat{x} = (A^T A + \hat{\alpha} I)^{-1} A^T b$, where $\hat{\alpha}$ is the unique positive root of the secular equation $\mathcal{G}(\alpha) = 0$.*

Second case: b belongs to the column span of A.

1. *If $\eta \geq \tau_2$, then the unique solution is $\hat{x} = 0$.*
2. *If $\tau_1 < \eta < \tau_2$, then the unique solution is $\hat{x} = (A^T A + \hat{\alpha} I)^{-1} A^T b$, where $\hat{\alpha}$ is the unique positive root of the secular equation $\mathcal{G}(\alpha) = 0$.*
3. *If $\eta \leq \tau_1$, then the unique solution is $\hat{x} = V \Sigma^{-1} b_1 = A^{\dagger} b$.*
4. *If $\eta = \tau_1 = \tau_2$, then there are infinitely many solutions that are given by $\hat{x} = \beta V \Sigma^{-1} b_1 = \beta A^{\dagger} b$, for any $0 \leq \beta \leq 1$.*

The above solution is suitable when the computation of the SVD of A is feasible. For large sparse matrices A, it is better to reformulate the secular equation as follows. Squaring both sides of (3.3) we obtain

$$\|(A^T A + \alpha I)^{-1} A^T b\|^2 \, \alpha^2 = \eta^2 \|A(A^T A + \alpha I)^{-1} A^T b \; - \; b\|^2 \, . \tag{3.22}$$

After some manipulation, we are led to

$$d^T (C + \alpha I)^{-2} d = \frac{\eta^2}{\alpha^2} \, \left[b^T b - d^T (C + \alpha I)^{-1} d - \alpha d^T (C + \alpha I)^{-2} d \right] \, ,$$

where we have defined $C = A^T A$ and $d = A^T b$. Therefore, finding α reduces to finding the positive root of

$$\begin{aligned} \mathcal{H}(\alpha) \triangleq \; & d^T (C + \alpha I)^{-2} d \\ & - \frac{\eta^2}{\alpha^2} \, \left[b^T b - d^T (C + \alpha I)^{-1} d - \alpha d^T (C + \alpha I)^{-2} d \right] \, . \end{aligned} \tag{3.23}$$

In this form, one can consider techniques similar to those suggested in [5] to find α efficiently.

4. Restricted perturbations. We have so far considered the case in which all the columns of the A matrix are subject to perturbations. It may happen in practice, however, that only selected columns are uncertain, while the remaining columns are known precisely. This situation can be handled by the approach of this paper, as we now clarify.

Given $A \in \mathbf{R}^{m \times n}$, we partition it into block columns,

$$A = \left[\begin{array}{cc} A_1 & A_2 \end{array} \right] \, ,$$

and assume, without loss of generality, that only the columns of A_2 are subject to perturbations while the columns of A_1 are known exactly. We then pose the following min-max problem.

Given $A \in \mathbf{R}^{m \times n}$, with $m \geq n$ and A full rank, $b \in \mathbf{R}^m$, and nonnegative real numbers (η_2, η_b), determine $\hat{x}$ such that

$$\min_{\hat{x}} \max \left\{ \| \left[\begin{array}{cc} A_1 & A_2 + \delta A_2 \end{array} \right] \hat{x} - (b + \delta b) \|_2 : \; \|\delta A_2\|_2 \leq \eta_2, \; \|\delta b\|_2 \leq \eta_b \right\}. \tag{4.1}$$

If we partition $\hat{x}$ accordingly with A_1 and A_2, say

$$\hat{x} = \left[\begin{array}{c} \hat{x}_1 \\ \hat{x}_2 \end{array} \right] \, ,$$

then we can write

$$\| \left[\begin{array}{cc} A_1 & A_2 + \delta A_2 \end{array} \right] \hat{x} - (b + \delta b) \|_2 \;=\; \|(A_2 + \delta A_2)\hat{x}_2 - (b - A_1\hat{x}_1 + \delta b)\|_2 \; .$$

Therefore, following the argument at the beginning of section 3, we conclude that the maximum over $(\delta A_2, \delta b)$ is achievable and is equal to

$$\|A_2\hat{x}_2 - (b - A_1\hat{x}_1)\|_2 \;+\; \eta_2\|\hat{x}_2\|_2 \;+\; \eta_b \; .$$

In this way, statement (4.1) reduces to the minimization problem

$$\min_{\hat{x}_1, \hat{x}_2} \left(\left\| \left[\begin{array}{cc} A_1 & A_2 \end{array} \right] \left[\begin{array}{c} \hat{x}_1 \\ \hat{x}_2 \end{array} \right] - b \right\|_2 \;+\; \eta_2\|\hat{x}_2\|_2 \;+\; \eta_b \right) \; . \tag{4.2}$$

This statement can be further reduced to the problem treated in Theorem 3.6 as follows. Introduce the QR decomposition of A, say

$$A = QR \;=\; Q \left[\begin{array}{cc} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & 0 \end{array} \right] \; ,$$

where we have partitioned R accordingly with the sizes of A_1 and A_2. Define

$$\left[\begin{array}{c} \bar{b}_{1A} \\ \bar{b}_{2A} \\ \bar{b}_2 \end{array} \right] \;=\; Q^T b \; .$$

Then (4.2) is equivalent to

$$\min_{\hat{x}_1, \hat{x}_2} \left(\left\| \left[\begin{array}{cc} R_{11} & R_{12} \\ 0 & R_{22} \\ 0 & 0 \end{array} \right] \left[\begin{array}{c} \hat{x}_1 \\ \hat{x}_2 \end{array} \right] - \left[\begin{array}{c} \bar{b}_{1A} \\ \bar{b}_{2A} \\ \bar{b}_2 \end{array} \right] \right\|_2 \;+\; \eta_2\|\hat{x}_2\|_2 \;+\; \eta_b \right) \; , \tag{4.3}$$

which can be further rewritten as

$$\min_{\hat{x}_1, \hat{x}_2} \left(\left\| \left[\begin{array}{c} R_{11}\hat{x}_1 + R_{12}\hat{x}_2 - \bar{b}_{1A} \\ R_{22}\hat{x}_2 - \bar{b}_{2A} \\ \bar{b}_2 \end{array} \right] \right\|_2 \;+\; \eta_2\|\hat{x}_2\|_2 \;+\; \eta_b \right) \; . \tag{4.4}$$

This shows that once the optimal $\hat{x}_2$ has been determined, the optimal choice for $\hat{x}_1$ is necessarily the one that annihilates the entry $R_{11}\hat{x}_1 + R_{12}\hat{x}_2 - \bar{b}_{1A}$. That is,

$$\hat{x}_1 = R_{11}^{-1} \left[\bar{b}_{1A} - R_{12}\hat{x}_2 \right] \; . \tag{4.5}$$

The optimal $\hat{x}_2$ is the solution of

$$\min_{\hat{x}_2} \left(\left\| \left[\begin{array}{c} R_{22} \\ 0 \end{array} \right] \hat{x}_2 - \left[\begin{array}{c} \bar{b}_{2A} \\ \bar{b}_2 \end{array} \right] \right\|_2 \;+\; \eta_2\|\hat{x}_2\|_2 \;+\; \eta_b \right) \; . \tag{4.6}$$

This optimization is of the same form as the problem stated earlier in Lemma 3.1 with $\hat{x}$ replaced by $\hat{x}_2$, η replaced by η_2, A replaced by $\left[\begin{array}{c} R_{22} \\ 0 \end{array} \right]$, and b replaced by $\left[\begin{array}{c} \bar{b}_{2A} \\ \bar{b}_2 \end{array} \right]$.

Therefore, the optimal $\hat{x}_2$ can be obtained by applying the result of Theorem 3.6. Once $\hat{x}_2$ has been determined, the corresponding $\hat{x}_1$ follows from (4.5).

5. Conclusion. In this paper we have proposed a new formulation for parameter estimation in the presence of data uncertainties. The problem incorporates a priori bounds on the size of the perturbations and admits a nice geometric interpretation. It also has a closed form solution that is obtained by solving a regularized least-squares problem with a regression parameter that is the unique positive root of a secular equation.

Several other interesting issues remain to be addressed. Among these, we state the following:

1. A study of the statistical properties of the min-max solution is valuable for a better understanding of its performance in stochastic settings.
2. The numerical properties of the algorithm proposed in this paper need also be addressed.
3. Extensions of the algorithm to deal with perturbations in submatrices of A are of interest and will be studied elsewhere.

We can also extend the approach of this paper to other variations that include uncertainties in a weighting matrix, multiplicatives uncertainties, etc. (see, e.g., [15]).

REFERENCES

[1] L. E. Ghaoui and H. Lebret, *Robust solutions to least-squares problems with uncertain data*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1035–1064.

[2] G. H. Golub, *Some modified matrix eigenvalue problems*, SIAM Rev., 15 (1973), pp. 318–344.

[3] 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.

[4] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1997.

[5] G. H. Golub and U. von Matt, *Generalized Cross-Validation for Large Scale Problems*, Tech. report SCCM-96-06, Computer Science Department, Stanford University, Stanford, CA, 1996.

[6] B. Hassibi, A. H. Sayed, and T. Kailath, *Linear estimation in Krein spaces - Part* I*: Theory*, IEEE Trans. Automat. Control, 41 (1996), pp. 18–33.

[7] S. V. Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, PA, 1991.

[8] S. M. Kay, *Fundamentals of Statistical Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1993.

[9] P. Khargonekar and K. M. Nagpal, *Filtering and smoothing in an* $H^\infty-$ *setting*, IEEE Trans. Automat. Control, AC-36 (1991), pp. 151–166.

[10] C. L. Lawson and R. J. Hanson, *Solving Least-Squares Problems*, SIAM, Philadelphia, PA, 1995. Revised republication of work first published by Prentice–Hall, Englewood Cliffs, NJ, 1974.

[11] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.

[12] A. H. Sayed, B. Hassibi, and T. Kailath, *Fundamental inertia conditions for the minimization of quadratic forms in indefinite metric spaces*, in Recent Developments in Operator Theory and Its Applications, Oper. Theory Adv. Appl. 87, Birkhäuser, Basel, 1996, pp. 309–347.

[13] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*, Addison–Wesley, Reading, MA, 1991.

[14] U. Shaked and Y. Theodor, $H^\infty-$*optimal estimation: A tutorial*, in Proc. IEEE Conference on Decision and Control, Tucson, AZ, December 1992, IEEE Computer Society Press, Los Alamitos, CA, pp. 2278–2286.

[15] S. Chandrasekaran, G. Golub, M. Gu, and A. H. Sayed, *Parameter estimation in the presence of bounded modeling errors*, IEEE Signal Processing Letters, 4 (1997), pp. 195–197.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 253–276, January 1998

017

A GENERALIZED HILBERT MATRIX PROBLEM AND CONFLUENT CHEBYSHEV–VANDERMONDE SYSTEMS*

HAO LU†

Abstract. This paper deals with fast solution of the generalized Hilbert matrix problem and confluent Chebyshev–Vandermonde systems. First, two methods for the generalized Hilbert matrix problem are presented. One is for the case where the points involved in the generalized Hilbert matrices satisfy a TH-relation introduced in the present paper, which include equidistant points, clustered points, and Chebyshev points. The approach is based on an $O(n \log n)$ fast multiplication of a Toeplitz plus Hankel matrix with a vector. The other method is to reduce the generalized Hilbert matrix problem to products of confluent Vandermonde-like matrices and dual confluent Vandermonde-like matrices with vectors by using J-matches and links of polynomials. Second, two strategies for confluent Chebyshev–Vandermonde systems are considered. Based on the result of the solution of confluent Vandermonde-like systems, the solution of confluent Chebyshev–Vandermonde systems for Chebyshev σ-points, i.e., the zeros of $T(\lambda) - \sigma$ with $|\sigma| < 1$, where $T(\lambda)$ is the Chebyshev polynomial of the first kind, is reduced to fast Fourier transforms (FFT) or to sine or cosine transforms by special choices of J-matches and links of Chebyshev polynomials, and hence we obtain some $O(n \log n)$ algorithms for the systems. The solution of Chebyshev–Vandermonde systems is also reduced to the generalized Hilbert matrix problem by using J-matches, links of Chebyshev polynomials, and the inversion of a class of generalized Hilbert matrices. This yields an $O(n \log n)$ algorithm for Chebyshev–Vandermonde systems for another class of practical points. Third, the results obtained are applied to related problems, for example, confluent Chebyshev–Vandermonde systems for near Chebyshev σ-points, Hermite interpolation in terms of Chebyshev polynomials, and a class of generalized Hilbert systems. Finally, numerical examples show quite accurate results even for large systems of equations.

Key words. generalized Hilbert matrix, confluent Vandermonde-like system, J-match and link of polynomials, Chebyshev polynomial, confluent Chebyshev–Vandermonde systems, Hermite interpolation, FFT, sine and cosine transform

AMS subject classifications. 65F05, 65F30, 68C25

PII. S0895479896307221

1. Introduction.

Let $H_p(\mathbf{t}, \mathbf{r})$ be an $n \times n$ generalized Hilbert matrix with the (k, j) entry

$$(H_p(\mathbf{t}, \mathbf{r}))_{kj} = \begin{cases} 1/(t_k - r_j)^p, & k \neq j, \\ 1/(t_k - r_k)^p, & t_k \neq r_k, \\ 0, & t_k = r_k, \end{cases} \tag{1}$$

where p is a positive integer, and t_k and r_k are points in the complex plane satisfying $t_k \neq t_j$, $r_k \neq r_j$, $t_k \neq s_j$ for $k \neq j$, $k, j = 1, 2, \ldots, n$. For the case $p = 1$ and $r_k = t_k$, denote $G = H_1(\mathbf{t}, \mathbf{t})$. Let $\mathbf{b}$ be any n-vector. In 1985, Golub [11] posed Trummer's problem as follows.

> Give an algorithm for computing $G\mathbf{b}$ in less than $O(n^2)$ multiplications. If this is impossible, show that it cannot be done.

The generalized Hilbert matrix problem is a generalization of Trummer's problem defined by the computation of the multiplications of generalized Hilbert matrices

* Received by the editors July 3, 1996; accepted for publication (in revised form) by N. J. Higham February 21, 1997.
http://www.siam.org/journals/simax/19-1/30722.html
† Department of Applied Mathematics, University of Twente, 7500 AE Enschede, The Netherlands (na.hlu@na-net.ornl.gov).

with vectors

$$H_1(\mathbf{t},\mathbf{r})\mathbf{b},\ H_2(\mathbf{t},\mathbf{r})\mathbf{b},\ \ldots,\ H_p(\mathbf{t},\mathbf{r})\mathbf{b}. \tag{2}$$

The problem is also called generalized Trummer's problem in [24]. Various applications of the problem can be found in the computation of conformal mappings [30], the numerical evaluation of singular integrals [8], [27], and particle simulations [12], [28].

In 1987, Gerasoulis, Grigoriadis, and Sun proposed an $O(n\log^2 n)$ algorithm for Trummer's problem, henceforth called the GGS algorithm [9]. Extending the GGS algorithm to include the matrices defined by (1) in the case $t_k \neq r_j$, $k,j = 1,2,\ldots,n$, Gerasoulis showed the existence of a fast algorithm with $O(n\log^2 n)$ time complexity for the multiplications of generalized Hilbert matrices with vectors $H_1(\mathbf{t},\mathbf{r})\mathbf{b}$, $H_2(\mathbf{t},\mathbf{r})\mathbf{b}$ [8]. For the case $t_k \neq r_j$, $k,j = 1,2,\ldots,n$, for any positive integer p, an $O(np\log n\log np)$ algorithm for the multiplication $H_p(\mathbf{t},\mathbf{r})\mathbf{b}$ was given by Lu in 1990 [22]. Based on the divide and conquer method and the fast polynomial arithmetic, an $O(np\log n\log\frac{n}{p})$ asymptotically fast algorithm for the generalized Hilbert matrix problem and an $O(np\log np + n\log^2 n)$ algorithm for computing multiplication $H_p(\mathbf{t},\mathbf{r})\mathbf{b}$ in the case $t_k \neq r_j$, $k,j = 1,2,\ldots,n$ were recently derived in [24]. For some special points we have much better results. Gerasoulis [8] derived an $O(n\log n)$ algorithm for $H_1(\mathbf{t},\mathbf{r})\mathbf{b}$ if $t_k = \cos((2k-1)\pi/2n)$ and $r_k = \cos(k\pi/n)$, $k = 1,2,\ldots,n$. For the original Trummer's problem, Reichel presented an $O(n\log n)$ algorithm for the case when t_k are equidistant points on a circle [27]. He proposed an approximation for the problem if there is a smooth 2π-periodic bijective function t such that $t_j = t(2\pi(j-1)/n)$, $j = 1,2,\ldots,n$.

Let f, g be functions defined on $S \subset \mathbb{C}$: $\to \mathbb{C}$ and $k \in S$, $t_k = f(k)$ and $r_k = g(k)$ for $k = 1,2,\ldots,n$. If there are functions d_1, d_2, q_1, q_2 on S such that $f(x) - g(y) = d_1(x)d_2(y)q_1(x+y)q_2(x-y)$, we say that points $\{t_1,t_2,\ldots,t_n\}$ and $\{r_1,r_2,\ldots,r_n\}$ have a TH-relation. It is straightforward to show that equidistant points, clustered points, and Chebyshev points satisfy a TH-relation. The first part of this paper deals with fast computation of the generalized Hilbert matrix problem for various point distributions. Two methods are proposed. The first one deals with the case where t_k and r_k satisfy a TH-relation. The generalized Hilbert matrix problem is reduced to p multiplications of Toeplitz plus Hankel matrices with vectors. This yields an $O(pn(p+\log n))$ algorithm for the generalized Hilbert matrix problem (2) and an $O(pn\log n)$ for the product $H_p(\mathbf{t},\mathbf{r})\mathbf{b}$. The approach is based on an $O(n\log n)$ algorithm for the multiplication of a Toeplitz plus Hankel matrix with a vector given in the present paper. The second method deals with the generalized Hilbert matrix problem for the case where either t_k or r_k are Chebyshev σ-points, i.e., zeros of $T(\lambda) - \sigma$, where σ is a constant satisfying $|\sigma| < 1$ and $T(\lambda)$ is the Chebyshev polynomial of the first kind. For the case $p \ll n$, for example, $p = 1,2$, or 3, the method yields an $O(n\log n)$ algorithm for the generalized Hilbert matrix problem.

Let $t_1,\ldots,t_p$ be p complex numbers, $n_1,\ldots,n_p$ be p positive integers, and $\mathbf{p}(\lambda) = (p_1(\lambda),\ldots,p_n(\lambda))^T$, where $n = \sum_{i=1}^{p} n_i$, and $P(\lambda) = \{p_1(\lambda),p_2(\lambda),\ldots,p_n(\lambda)\}$ 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 [17]), denoted by $V_c(\mathbf{p})$, is given by

$$V_c(\mathbf{p}) = (B_1,B_2,\ldots,B_p), \tag{3}$$

where B_k is an $n \times n_k$ matrix with the (i,j) entry $\big(p_i(\lambda)\big)^{(j-1)}\big|_{\lambda=t_k}$. In the case of $n_1 = n_2 = \cdots = n_p = 1$, $V_c(\mathbf{p})$ yields a Vandermonde-like matrix [7], [16]. We denote

the Vandermonde-like matrix by $V(\mathbf{p})$. Consider confluent Vandermonde-like systems

$$V_c(\mathbf{p})\mathbf{x} = \mathbf{b}. \tag{4}$$

These systems are associated with the construction of quadrature formulae [1],[13], [20], [26] and the approximation of linear functionals [2], [29]. Study of fast solution of Vandermonde systems started with some $O(n^2)$ algorithms in the early 1970s by Björck, Elfving, and Pereyra [3], [4]. Many $O(n^2)$ algorithms for Vandermonde systems, Vandermonde-like systems, and confluent Vandermonde-like systems are available in the literature. See the historical surveys in [19] or [25] for further details. Error analysis for some algorithms is presented by Higham in [15], [17]. The $O(n^2)$ complexity of algorithms for the systems is not optimal. $O(n\log^2 n)$ algorithms are derived by Lu [21], [23], [24], and [25] based on J-matches and links of polynomials.

If $p_k(\lambda)$ are Chebyshev polynomials, then $V_c(\mathbf{P})$ is called the confluent Chebyshev–Vandermonde matrix. The second part of this paper focuses on the practicality of the approach in [25] for confluent Chebyshev–Vandermonde systems for various distributions of points t_i. In section 4, we present $O(n\log n)$ algorithms for confluent Chebyshev–Vandermonde linear systems for Chebyshev σ-points by using the result of confluent Vandermonde-like systems given in [25] with special choices of J-matches and links for Chebyshev polynomials. In section 5, relations between the solution of Chebyshev–Vandermonde systems and the generalized Hilbert matrix problem are discussed by using J-matches and links of Chebyshev polynomials based on the inversion of a class of generalized Hilbert matrices [6]. The solution of Chebyshev–Vandermonde systems is reduced to the generalized Hilbert matrix problem. For the case where $t_k = \cos\frac{(k-1)\pi+\alpha}{n}$ for Chebyshev–Vandermonde systems, we obtain an $O(n\log n)$ algorithm by using the result in section 3 for the generalized Hilbert matrix problem. In section 6, we consider applications of the results in the previous sections to some related problems: (a) solution of confluent Chebyshev–Vandermonde systems for near Chebyshev σ-points, i.e., most points of $t_1, t_2, \ldots, t_p$ are zeros of $T_p(\lambda) - \sigma$; (b) dual confluent Chebyshev–Vandermonde systems, i.e., Hermite interpolation of polynomials in terms of Chebyshev polynomials, for Chebyshev σ-points or $t_k = \cos\frac{((k-1)+\alpha)\pi}{n}$; and (c) solution of generalized Hilbert systems $H_1(\mathbf{r},\mathbf{t})\mathbf{x} = \mathbf{b}$ if t_k and r_j satisfy a TH-relation. An $O(n\log n)$ algorithm is derived for the generalized Hilbert systems. Numerical examples are presented in section 7. Finally, conclusions are made in section 8.

2. Fast computation for the generalized Hilbert matrix problem I. The aim of this section is to construct a fast algorithm for the generalized Hilbert matrix problem for the case where the points t_i and r_j satisfy a TH-relation. The approach is based on the product of a Toeplitz plus Hankel matrix with a vector.

An $n\times n$ matrix T is a Toeplitz matrix if the (i,j) entry of T satisfies $(T)_{ij} = t_{j-i}$. An $n \times n$ circulant matrix C is a special Toeplitz matrix satisfying $(C)_{ij} = c_{j-i}$ and $c_{-k} = c_{n-k}$ for $k = 1, \ldots, n-1$, or briefly $C = \mathrm{circ}(c_0, c_1, \ldots, c_{n-1})$. Circulant matrices are diagonalized by the Fourier matrix F_n (see Davis [5, Theorem 3.2.2]), i.e.,

$$C = F_n \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{n-1}) F_n^*, \tag{5}$$

where $(F_n)_{kj} = \omega_n^{(k-1)(j-1)}/\sqrt{n}$, $\omega_n = \exp(-2\pi i/n)$ is the primitive nth root of unity and $\lambda_k = \sum_{j=0}^{n-1} c_j \omega_n^{kj}$. An $n \times n$ Hankel matrix H is a matrix with the (i,j)

entry $(H)_{ij} = h_{i+j-2}$. An $n \times n$ skew-circulant matrix S is a special Hankel matrix satisfying $(S)_{ij} = s_{i+j-2}$ and $s_{n+k} = s_k$ for $k = 0, \ldots, n-2$, or briefly, $S = \mathrm{scirc}(s_0, s_1, \ldots, s_{n-1})$. For a Toeplitz matrix T denote $T(k) = t_k$, $k = 0, \pm 1, \ldots, \pm(n-1)$ and for a Hankel matrix H denote $H(k) = h_k$, $k = 0, 1, \ldots, 2n-2$. It is straightforward to see that the product $T\mathbf{b}$ can be computed by embedding T into a $2n \times 2n$ circulant matrix. Therefore, the multiplication $T\mathbf{b}$ is computed by an FFT of order $2n$ three times as well as the multiplication of a Hankel matrix with a vector. This yields an $O(n \log n)$ algorithm for the multiplication of a Toeplitz plus Hankel matrix with a vector. To further reduce the arithmetic operations, we now present an algorithm by using an FFT of order $2n$ four times instead of six. To this end, we need the following property of skew-circulant matrices.

LEMMA 2.1. *Let* $S = \mathrm{scirc}(a_0, a_1, \ldots, a_{n-1})$. *Then*

$$S = F_n \Lambda F_n^*,$$

where $(F_n)_{kj} = \omega_n^{(k-1)(j-1)}/\sqrt{n}$, $k, j = 1, 2, \ldots, n$, $\omega_n = \exp(-2\pi i/n)$ *is the primitive* n*th root of unity and*

$$\Lambda = \mathrm{qdiag}(\lambda_0, \lambda_1, \ldots, \lambda_{n-1}) \equiv \begin{pmatrix} \lambda_0 & 0 & \ldots & 0 & 0 \\ 0 & \ldots & 0 & 0 & \lambda_{n-1} \\ \ldots & 0 & 0 & \lambda_{n-2} & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \lambda_2 & 0 & \ldots \\ 0 & \lambda_1 & 0 & \ldots & 0 \end{pmatrix}$$

with $\lambda_k = \sum_{j=0}^{n-1} a_j \omega^{kj}$, $k = 0, 1, \ldots, n-1$.

Proof. Let $E = \mathrm{scirc}(1, 0, \ldots, 0)$. An elementary computation shows $E^{-1} = E$. Since $ES = \mathrm{circ}(a_0, a_1, \ldots, a_{n-1})$, applying (5) shows that

$$ES = F_n \mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{n-1}) F_n^*.$$

Therefore, $S = EF_nEE\mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{n-1})F_n^*$. It is straightforward to check that $EF_nE = F_n$ and $E\mathrm{diag}(\lambda_0, \lambda_1, \ldots, \lambda_{n-1}) = \Lambda$, which imply the desired result. ∎

Now we embed an $n \times n$ Toeplitz matrix T into a $2n \times 2n$ circulant matrix $C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$ with $C_{11} = T$ and embed an $n \times n$ Hankel matrix H into a $2n \times 2n$ skew-circulant matrix $S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$ with $S_{11} = H$. Let $\widetilde{\mathbf{b}} = \begin{pmatrix} \mathbf{b} \\ 0 \end{pmatrix}$. We then obtain $(T+H)\mathbf{b}$ from $(C+S)\widetilde{\mathbf{b}}$. Denote $C = \mathrm{circ}(c_0, c_1, \ldots, c_{2n-1})$ and $S = \mathrm{scirc}(s_0, s_1, \ldots, s_{2n-1})$. It follows from (5) and Lemma 2.1 that $C+S = F_{2n}\Gamma F_{2n}^*$, where $\Gamma = \mathrm{diag}(\mu_0, \mu_1, \ldots, \mu_{2n-1}) + \mathrm{qdiag}(\beta_0, \beta_1, \ldots, \beta_{2n-1})$ and

$$\mu_k = \sum_{j=0}^{2n-1} c_j \omega_{2n}^{kj}, \qquad \beta_k = \sum_{j=0}^{2n-1} s_j \omega_{2n}^{kj}. \tag{6}$$

Hence, we compute $(T+H)\mathbf{b}$ as follows.

ALGORITHM 1. Let T and H be $n \times n$ Toeplitz and Hankel matrices, respectively, and let $\mathbf{b}$ be a vector. This algorithm computes $(T+H)\mathbf{b}$. The result overwrites $\mathbf{b}$.

1. $\mathbf{b}(1:n) = \mathbf{b}$, $\mathbf{b}(n+1:2n) = \mathrm{zeros}(1, n)$, $\mathbf{h} = (H(0), H(1), \ldots, H(2n-2), 0)$, $\mathbf{t} = (T(0), \ldots, T(n-1), 0, T(1-n), \ldots, T(-1))$.
2. $\mathbf{b} = \mathrm{ifft}(\mathbf{b})$; $\mathbf{t} = \mathrm{fft}(\mathbf{t})$; $\mathbf{h} = \mathrm{fft}(\mathbf{h})$.

3. $\mathbf{b}(1) = (\mathbf{t}(1) + \mathbf{h}(1)) * \mathbf{b}(1)$.
   $\mathbf{b}(2:2n) = \mathbf{t}(2:2n).*\mathbf{b}(2:2n) + \mathbf{h}(2n:-1:2).*\mathbf{b}(2n:-1:2)$.
4. $\mathbf{b}$=fft($\mathbf{b}$).
5. $\mathbf{b}$=$\mathbf{b}$(1:n).

The dominant computation of the algorithm is to perform an FFT of order $2n$ four times, and hence the algorithm needs $O(n \log n)$ operations.

DEFINITION 2.2. *Let f, g be functions defined on $S \subset \mathbb{C}$: $\to \mathbb{C}$. If there are functions d_1, d_2, q_1, q_2 on S such that*

$$f(x) - g(y) = d_1(x)d_2(y)q_1(x+y)q_2(x-y), \tag{7}$$

we call that f and g satisfy a TH-relation on S. If $k \in S$, $t_k = f(k)$, and $r_k = g(k)$ for $k = 1, \ldots, n$, we say that points $\{t_1, t_2, \ldots, t_n\}$ and $\{r_1, r_2, \ldots, r_n\}$ have a TH-relation.

We consider the generalized Hilbert matrix problem (2) if t_k and r_k satisfy a TH-relation. It is straightforward to check that the following pairs satisfy TH-relations.

1. $f(x) = ax + b$, $g(x) = ax + c$.
2. $f(x) = (ax + b)^2 + d$, $g(x) = (ax + c)^2 + d$.
3. $f(x) = c\cos(ax + b)$, $g(x) = c\cos(ax + d)$.
4. $f(x) = c\sin(ax + b)$, $g(x) = c\sin(ax + d)$.
5. $f(x) = c\cos(ax + b)$, $g(x) = c\sin(ax + d)$.
6. $f(x) = c\tan(ax + b)$, $g(x) = c\tan(ax + d)$.
7. $f(x) = ab^x$, $g(x) = cb^x$.

Equidistant points, clustered points, and Chebyshev points can be obtained from the first three function pairs. It is not hard to find other function pairs satisfying TH-relations. For example, except for number 2, for the function pairs mentioned above the corresponding function pairs (f^2, g^2) have TH-relations.

PROPOSITION 2.3. *Assume that function pair (f, g) satisfies a TH-relation and all functions in equation (7) are differentiable. Then for $f(x) \neq g(y)$,*

$$\frac{f'(x)}{f(x) - g(y)} = \frac{q_1'(x+y)}{q_1(x+y)} + \frac{q_2'(x-y)}{q_2(x-y)} + \frac{d_1'(x)}{d_1(x)}, \tag{8}$$

$$\frac{g'(y)}{f(x) - g(y)} = \frac{q_1'(x+y)}{q_1(x+y)} - \frac{q_2'(x-y)}{q_2(x-y)} + \frac{d_2'(y)}{d_2(y)}. \tag{9}$$

Proof. Computing derivative for x in (7) we have

$$\begin{aligned} f'(x) = {} & d_1'(x)d_2(y)q_1(x+y)q_2(x-y) \\ & d_1(x)d_2(y)q_1'(x+y)q_2(x-y) + d_1(x)d_2(y)q_1(x+y)q_2'(x-y). \end{aligned}$$

Hence, (8) follows immediately. The equality (9) follows in a similar way. ∎

If $f'(x) \neq 0$, by induction based on (8) it is easy to show the fundamental formula

$$\frac{1}{(f(x) - g(y))^{m+1}} = \frac{(-1)^m}{m!} \left(\sum_{i=0}^{m} u_{mi}(x)(h^{(i)}(x+y) + t^{(i)}(x-y)) + v_m(x) \right) \tag{10}$$

for reducing the generalized Hilbert matrix problem to products of Toeplitz plus Hankel matrices with vectors, provided functions $q_1(x), q_2(x)$, and $d_1(x)$ satisfy the conditions required, where

$$u_{mm}(x) = u_{m-1,m-1}(x)/f'(x), \qquad u_{m0}(x) = u_{m-1,0}'(x)/f'(x),$$

$$u_{mi}(x) = (u_{m-1,i-1}(x) + u'_{m-1,i}(x))/f'(x), \qquad v_m = v'_{m-1}(x)/f'(x),$$
$$u_{00}(x) = 1/f'(x), \qquad v_0(x) = d'_1(x)/(d_1(x)f'(x)),$$
$$h(x) = q'_1(x)/q_1(x), \qquad t(x) = q'_2(x)/q_2(x).$$

Assume that the function pair (f, g) satisfies a TH-relation and $t_k = f(k)$, $r_j = g(j)$. Under the condition $r_k \neq t_j$ for $k \neq j$, $k, j = 1, 2, \ldots, n$ we find that $q_1(2k) \neq 0$ for $1 < k < n$. For convenience assume that $f'(k) \neq 0$ for $k = 1, 2, \ldots, n$. Define a Toeplitz matrix T_i by

$$T_i(0) = \begin{cases} 0, & \text{if } q_2(0) = 0, \\ t^{(i)}(0), & \text{otherwise,} \end{cases}$$

and $T_i(k) = t^{(i)}(k)$ for $|k| = 1, \ldots, n-1$, and define a Hankel matrix H_i by

$$H_i(2k-2) = \begin{cases} 0, & \text{if } q_1(2k) = 0 \text{ or } q_2(0) = 0, \\ h^{(i)}(2k), & \text{otherwise,} \end{cases}$$

for $k = 1, n$ and $H_i(k) = h^{(i)}(k+2)$ for $k = 1, \ldots, 2n-3$. It follows from (10) that

$$H_{m+1}(\mathbf{t}, \mathbf{r}) = \frac{(-1)^m}{m!} \left(\sum_{i=0}^{m} D_{mi}(H_i + T_i) + D_m J + G_m \right), \tag{11}$$

where J is the matrix with all entries 1,

$$D_{mi} = \operatorname{diag}(u_{mi}(1), \ldots, u_{mi}(n)), \quad i = 0, 1, \ldots, m,$$
$$D_m = \operatorname{diag}(v_m(1), \ldots, v_m(n)), \quad G_m = \operatorname{diag}(g_{m1}, \ldots, g_{mn}),$$
$$g_{mk} = \begin{cases} -\sum_{i=0}^{m} u_{mi}(k) H_i(2k-2), & q_2(0) = 0, \\ -\sum_{i=0}^{m} u_{mi}(k) T_i(0), & k = 1, n \text{ and } q_1(2k) = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, Algorithm 1 for the product of a Toeplitz plus Hankel matrix with a vector is efficient for the generalized Hilbert matrix problem (2). In practice, p is considerable small. If we ignore the computational cost for functions u_{mi}, v_m, $t^{(i)}$, $h^{(i)}$, $m = 1, \ldots, p$, $i = 0, 1, \ldots, m$, it is readily seen that the number of operations is $O(pn(p + \log n))$ for the generalized Hilbert matrix problem (2) and $O(pn \log n)$ for the product $H_p\mathbf{b}$. If $g'(x) \neq 0$, one can also derive a method for the generalized Hilbert matrix problem by using (9) if $q_1(x), q_2(x)$, and $d_2(x)$ satisfy the corresponding conditions required. An application of the approach in this section to a class of generalized Hilbert systems will be given in section 6.3.

3. Fast computation for the generalized Hilbert matrix problem II. In this section we consider a new method for the generalized Hilbert matrix problem (2) based on J-matches and links of polynomials. This new approach is more efficient for some practical cases, for example, t_k or r_k are Chebyshev σ-points.

Let $P(\lambda) = \{p_1(\lambda), \ldots, p_n(\lambda)\}$ be a basis of the linear space $\mathbb{C}_{n-1}[\lambda]$ of all complex polynomials of degree at most $n-1$, and $Q(\lambda) = \{q_1(\lambda), \ldots, q_n(\lambda)\}$ is a J-match of

$P(\lambda)$ with a link $p(\lambda)$, i.e.,

$$\frac{p(\lambda)-p(\mu)}{\lambda-\mu}=\sum_{i=1}^{n}p_i(\lambda)q_{n-i+1}(\mu). \tag{12}$$

See [25] for details. For any basis $\{p_1(\lambda),\dots,p_n(\lambda)\}$ of the space $\mathbb{C}_{n-1}[\lambda]$, the proof of Theorem 2.1 in [25] for the existence of a J-match shows, actually, the following stronger result than the theorem.

COROLLARY 3.1. *Let $P(\lambda)=\{p_1(\lambda),\dots,p_n(\lambda)\}$ be a basis of the linear space $\mathbb{C}_{n-1}[\lambda]$ of all complex polynomials of degree at most $n-1$ and let $p(\lambda)$ be a polynomial of degree n. Then there exists a unique J-match of $P(\lambda)$ with the link $p(\lambda)$.*

Calculating the mth derivatives for λ and μ in (12), respectively, shows that

$$\sum_{i=0}^{m-1}\frac{(-1)^i m!p^{(m-i)}(\lambda)}{(m-i)!(\lambda-\mu)^{i+1}}+(-1)^m m!\frac{p(\lambda)-p(\mu)}{(\lambda-\mu)^{m+1}}=\sum_{i=1}^{n}p_i^{(m)}(\lambda)q_{n-i+1}(\mu), \tag{13}$$

$$-\sum_{i=0}^{m-1}\frac{m!p^{(m-i)}(\mu)}{(m-i)!(\lambda-\mu)^{i+1}}+m!\frac{p(\lambda)-p(\mu)}{(\lambda-\mu)^{m+1}}=\sum_{i=1}^{n}p(\lambda)q_{n-i+1}^{(m)}(\mu), \tag{14}$$

which implies that if $p(\lambda)=p(\mu)$ for $\lambda\neq\mu$,

$$\sum_{i=0}^{m-1}\frac{(-1)^i(m+1)!p^{(m+1-i)}(\lambda)}{(m+1-i)!(\lambda-\mu)^{i+1}}+\frac{(-1)^m(m+1)!p'(\lambda)}{(\lambda-\mu)^{m+1}} =\sum_{i=1}^{n}p_i^{(m+1)}(\lambda)q_{n-i+1}(\mu), \tag{15}$$

$$-\sum_{i=0}^{m-1}\frac{(m+1)!p^{(m+1-i)}(\mu)}{(m+1-i)!(\lambda-\mu)^{i+1}}-!\frac{(m+1)!p'(\mu)}{(\lambda-\mu)^{m+1}}=\sum_{i=1}^{n}p_i(\lambda)q_{n-i+1}^{(m+1)}(\mu). \tag{16}$$

PROPOSITION 3.2. *Let p, p_i, q_i, $i=1,\dots,n$ be sufficiently smooth functions. If*

$$p(x)-p(y)=(x-y)\sum_{i=1}^{n}p_i(x)q_{n-i+1}(y), \tag{17}$$

then for a positive integer m,

$$\sum_{i=1}^{n}p_i^{(m-1)}(x)q_{n-i+1}(x)=p^{(m)}(x)/m, \tag{18}$$

$$\sum_{i=1}^{n}p_i(x)q_{n-i+1}^{(m-1)}(x)=p^{(m)}(x)/m. \tag{19}$$

Proof. By computing the mth derivative for x in (17) and then replacing y by x, (18) follows immediately. The equality (19) follows in a similar way. ∎

Let $p(\lambda)$ be a link of $\{P(\lambda),Q(\lambda)\}$. It follows from Corollary 2.2 in [25] that $\deg(p(\lambda))=n$. Assume $p(r_i)=\nu$, where ν is a constant. Then $p(t_i)\neq\nu$ if $t_i\neq r_i$, and $p'(t_i)\neq 0$ if $t_i=r_i$ simply because the polynomial $p(\lambda)-\nu$ of degree n has just n distinct zeros $r_1,r_2,\dots,r_n$. Let

$$V(\mathbf{q})=\begin{pmatrix} q_1(r_1) & q_1(r_2) & \cdots & q_1(r_n)\\ q_1(r_2) & q_2(r_2) & \cdots & q_2(r_n)\\ \cdots & \cdots & \cdots & \cdots\\ q_n(r_1) & q_n(r_2) & \cdots & q_n(r_n)\end{pmatrix}$$

and $\widetilde{V}^{(m)}(\mathbf{p}) = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)^T$, where

$$\mathbf{v}_k = \begin{cases} (p_n^{(m)}(t_k), \ldots, p_2^{(m)}(t_k), p_1^{(m)}(t_k))^T, & \text{if } t_k \neq r_k, \\ (p_n^{(m+1)}(t_k), \ldots, p_2^{(m+1)}(t_k), p_1^{(m+1)}(t_k))^T, & \text{if } t_k = r_k. \end{cases}$$

Applying (13), (15), and (18) shows that

$$H_{m+1}(\mathbf{t}, \mathbf{r}) = (-1)^m D_1 \left(\frac{1}{m!} D_2 (\widetilde{V}^{(m)}(\mathbf{p}) V(\mathbf{q}) - D_3) - \sum_{i=0}^{m-1} D_{mi} H_{i+1}(\mathbf{t}, \mathbf{r}) \right), \tag{20}$$

where $D_1 = \operatorname{diag}(a_1, a_2, \ldots, a_n)$, $D_2 = \operatorname{diag}(b_1, b_2, \ldots, b_n)$, $D_3 = \operatorname{diag}(c_1, c_2, \ldots, c_n)$, and $D_{mi} = \frac{(-1)^i}{(m-i)!} \operatorname{diag}(d_{mi}(1), d_{mi}(2), \ldots, d_{mi}(n))$,

$$a_k = \begin{cases} 1/(p(t_k) - \nu), & \text{if } t_k \neq r_k, \\ 1/p'(t_k), & \text{if } t_k = r_k, \end{cases} \qquad b_k = \begin{cases} 1, & \text{if } t_k \neq r_k, \\ 1/(m+1), & \text{if } t_k = r_k, \end{cases}$$

$$c_k = \begin{cases} 0, & \text{if } t_k \neq r_k, \\ \dfrac{p^{(m+2)}(t_k)}{m+2}, & \text{if } t_k = r_k, \end{cases} \qquad d_{mi}(k) = \begin{cases} p^{(m-i)}(t_k), & \text{if } t_k \neq r_k, \\ \dfrac{p^{(m+1-i)}(t_k)}{m+1-i}, & \text{if } t_k = r_k. \end{cases}$$

If $p(t_k) = \nu$, $k = 1, 2, \ldots, n$, denote

$$\widetilde{V}(\mathbf{p}) = \begin{pmatrix} p_n(t_1) & \cdots & p_2(t_1) & p_1(t_1) \\ p_n(t_2) & \cdots & p_2(t_2) & p_1(t_2) \\ \cdots & \cdots & \cdots & \cdots \\ p_n(t_n) & \cdots & p_2(t_n) & p_1(t_n) \end{pmatrix}$$

and $V^{(m)}(\mathbf{q}) = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$, where

$$\mathbf{u}_k = \begin{cases} (q_0^{(m)}(r_k), q_1^{(m)}(r_k), \ldots, q_{n-1}^{(m)}(r_k))^T, & \text{if } r_k \neq t_k, \\ (q_0^{(m+1)}(r_k), q_1^{(m+1)}(r_k), \ldots, q_{n-1}^{(m+1)}(r_k))^T, & \text{if } r_k = t_k. \end{cases}$$

Similarly, using (14), (16), and (19) we formulate

$$H_{m+1}(\mathbf{t}, \mathbf{r}) = \left(\sum_{i=0}^{m-1} H_{i+1} \widetilde{D}_{mi} + \frac{1}{m!} (\widetilde{V}(\mathbf{p}) V^{(m)}(\mathbf{p}) - D_3) D_2 \right) \widetilde{D}_1, \tag{21}$$

where $\widetilde{D}_1 = \operatorname{diag}(\widetilde{a}_1, \widetilde{a}_2, \ldots, \widetilde{a}_n)$ and $\widetilde{D}_{mi} = \frac{1}{(m-i)!} \operatorname{diag}(\widetilde{d}_{mi}(1), \widetilde{d}_{mi}(2), \ldots, \widetilde{d}_{mi}(n))$,

$$\widetilde{a}_k = \begin{cases} 1/(\nu - p(r_k)), & \text{if } r_k \neq t_k, \\ -1/p'(r_k), & \text{if } r_k = t_k, \end{cases} \qquad \widetilde{d}_{mi}(k) = \begin{cases} p^{(m-i)}(r_k), & \text{if } r_k \neq t_k, \\ \dfrac{p^{(m+1-i)}(r_k)}{m+1-i}, & \text{if } r_k = t_k. \end{cases}$$

If r_k or t_k are roots of $p(\lambda) - \nu$ and the arithmetic operations of matrix-vector products $\widetilde{V}^{(m)}(\mathbf{p}) V(\mathbf{q}) \mathbf{b}$ or $\widetilde{V}(\mathbf{p}) V^{(m)}(\mathbf{q}) \mathbf{b}$ do not exceed $C(n)$, using (20) or (21) we can compute the generalized Hilbert matrix problem by using $O(p(pn + C(n)))$ operations and matrix-vector product $H_p(\mathbf{t}, \mathbf{r}) \mathbf{b}$ by $O(p(n + C(n)))$ operations. If $p \ll n$, for example, $p = 1, 2$, or 3, consider the cases where (i) t_k are Chebyshev σ_1-points and r_k are Chebyshev σ_2-points; (ii) t_k are Chebyshev σ-points and $r_k = \cos\frac{(k-1)\pi+\alpha}{n}$; or (iii) $t_k = \cos\frac{(k-1)\pi+\alpha}{n}$ and r_k are Chebyshev σ-points. We choose

$p_k(\lambda) = T_{k-1}(\lambda)$ and $q_k(\lambda) = 2 * U_{k-1}(\lambda)$ for $k < n$ and $q_n(\lambda) = U_{n-1}(\lambda)$, where $T_k(\lambda) = \cos(k \arccos(\lambda))$ is the Chebyshev polynomial of the first kind of degree k and $U_k(\lambda) = \sin((k+1)\arccos(\lambda))/\sin(\arccos(\lambda))$ is the Chebyshev polynomial of the second kind of degree k. It follows from [10] or [25] that

$$\frac{T_n(\lambda) - T_n(\mu)}{\lambda - \mu} = 2\sum_{k=0}^{n-2} T_{n-k-1}(\lambda)U_k(\mu) + T_0(\lambda)U_{n-1}(\mu), \tag{22}$$

which implies that $\{q_1(\lambda), \ldots, q_n(\lambda)\}$ is the J-match of $\{p_1(\lambda), \ldots, p_n(\lambda)\}$ with the link $T_n(\lambda)$. For these three cases it is not hard to check that the corresponding $\widetilde{V}^{(m)}(\mathbf{p})V(\mathbf{q})\mathbf{b}$ and $\widetilde{V}(\mathbf{p})V^{(m)}(\mathbf{q})\mathbf{b}$ can be computed by FFT or sine or cosine transforms. Therefore, our approach yields an $O(n \log n)$ algorithm for the generalized Hilbert matrix problem. An application of this result to Chebyshev–Vandermonde systems will be given in section 5.

4. Solution of confluent Chebyshev–Vandermonde systems for Chebyshev σ-points. If $p_k(\lambda) = T_{k-1}(\lambda)$ are Chebyshev polynomials of the first kind, the matrix $V_c(\mathbf{p})$, denoted by $V_{c,t}$, is called the confluent Chebyshev–Vandermonde matrix of the first kind. If $p_k(\lambda) = U_{k-1}(\lambda)$ are Chebyshev polynomials of the second kind, the matrix $V_c(\mathbf{p})$, denoted by $V_{c,u}$, is called the confluent Chebyshev–Vandermonde matrix of the second kind. Based on the results of confluent Vandermonde-like systems given in [25], the solution of confluent Chebyshev–Vandermonde systems is represented by trigonometric functions. Furthermore, applications of FFT or sine or cosine transforms yield some $O(n \log n)$ algorithms for confluent Chebyshev–Vandermonde systems $V_{c,t}\mathbf{x} = \mathbf{b}$ and $V_{c,u}\mathbf{x} = \mathbf{b}$ for Chebyshev σ-points, i.e., the zeros of $T_p(\lambda) - \sigma$ with $|\sigma| < 1$, of the following three problems.

- Problem 1. $n_1 = n_2 = \cdots = n_p = 1$.
- Problem 2. $n_1 = n_2 = \cdots = n_p = 2$.
- Problem 3. $n_1 = n_2 = \cdots = n_p = 3$.

Let $A(\lambda)$ and $B(\lambda)$ be two polynomials. For convenience, $\mathrm{quot}(A(\lambda), B(\lambda))$ denotes the quotient of polynomial division $A(\lambda)/B(\lambda)$, i.e., ignoring the remainder $R(\lambda)$: $A(\lambda) = B(\lambda)\mathrm{quot}(A(\lambda), B(\lambda)) + R(\lambda)$. For readers' convenience, the result on the solution of confluent Vandermonde-like systems is stated in the following theorem.

THEOREM 4.1. *Let $V_c(\mathbf{p})$ be a confluent Vandermonde-like matrix defined by $\mathbf{p}(\lambda) = (p_1(\lambda), p_2(\lambda), \ldots, p_n(\lambda))^T$ via (3) with $t_i \neq t_j$, $i \neq j$, $i, j = 1, \ldots, p$, where $P(\lambda) = \{p_1(\lambda), p_2(\lambda), \ldots, p_n(\lambda)\}$ is a basis of $\mathbb{C}_{n-1}[\lambda]$, and*

$$r(\lambda) = \prod_{i=1}^{p}(\lambda - t_i)^{n_i}, \quad r_i(\lambda) = r(\lambda)/(\lambda - t_i)^{n_i}, \quad i = 1, \ldots, p.$$

Then the solution of the confluent Vandermonde-like systems (4) is given by

$$x_i = \frac{1}{(k-1)!(n_j - k)!}\left(\frac{v(\lambda)}{r_j(\lambda)}\right)^{(n_j-k)}\Bigg|_{\lambda=t_j}, \tag{23}$$

$$i = m_j + k, \quad 1 \le j \le p, \quad 1 \le k \le n_j, \quad m_1 = 0, \quad m_j = \sum_{t=1}^{j-1} n_t,$$

where $v(\lambda) = \mathrm{quot}(r(\lambda)b(\lambda), p(\lambda))$, and the polynomial $b(\lambda) = \sum_{i=1}^{n} b_i q_{n-i+1}(\lambda)$, and $Q(\lambda) = \{q_1(\lambda), \ldots, q_n(\lambda)\}$ is a J-match of $P(\lambda)$, and $p(\lambda)$ is a link of $\{P(\lambda), Q(\lambda)\}$.

If $Q(\lambda) = \{q_1(\lambda), \ldots, q_n(\lambda)\}$ is a J-match of $P(\lambda)$, $Q(\lambda)$ is occasionally called a J-match of the matrix $V_c(\mathbf{p})$ or the systems $V_c(\mathbf{p})\mathbf{x} = \mathbf{b}$.

4.1. Problem 1. Let $p(\lambda)$ be a link of $\{P(\lambda), Q(\lambda)\}$ and assume that there are n distinct zeros $t_1, t_2, \ldots, t_n$ of $p(\lambda)-\sigma$, where σ is a constant. Consider Vandermonde-like systems $V(\mathbf{p})\mathbf{x} = \mathbf{b}$. Then $p(\lambda) - \sigma = dr(\lambda)$, $v(\lambda) = d^{-1}b(\lambda)$, and $r_k(t_k) = d^{-1}p'(t_k)$, where d is a nonzero constant. Theorem 4.1 shows that the solution of the systems is given by

$$x_k = b(t_k)/p'(t_k). \tag{24}$$

The equality (22) implies that the basis $\{2U_0(\lambda), \ldots, 2U_{n-2}(\lambda), U_{n-1}(\lambda)\}$ and the basis $\{T_0(\lambda), 2T_1(\lambda), \ldots, 2T_{n-1}(\lambda)\}$ are J-matches of $\{T_0(\lambda), T_1(\lambda), \ldots, T_{n-1}(\lambda)\}$ and $\{U_0(\lambda), U_1(\lambda), \ldots, U_{n-1}(\lambda)\}$ with the link $T_n(\lambda)$, respectively. If t_k are zeros of $T_n(\lambda) - \sigma$ with $|\sigma| < 1$, then $t_k = \cos\frac{k-\alpha_k}{p}\pi$, where $\alpha_k = 1-\alpha$ if k is odd, $\alpha_k = \alpha$ if k is even, and $0 < \alpha < 1$ is a constant satisfying $\cos\alpha\pi = \sigma$. For the systems $V_{c,t}\mathbf{x} = \mathbf{b}$, by choosing the J-match $\{2U_0(\lambda), \ldots, 2U_{n-2}(\lambda), U_{n-1}(\lambda)\}$ with the link $T_n(\lambda)$ it follows from Theorem 4.1 that the solution of the linear systems $V_{c,t}\mathbf{x} = \mathbf{b}$ is given by

$$x_j = \frac{(-1)^{j-1}}{n\sin\alpha\pi}\left(b_1(-1)^{j-1}\sin\alpha\pi + 2\sum_{k=2}^{n} b_k \sin\frac{(n-k+1)(j-\alpha_j)\pi}{n}\right). \tag{25}$$

We choose the J-match $\{T_0(\lambda), 2T_1(\lambda), \ldots, 2T_{n-1}(\lambda)\}$ with the link $T_n(\lambda)$ for the linear systems $V_{c,u}\mathbf{x} = \mathbf{b}$. Applying Theorem 4.1 yields the solution

$$x_j = \frac{(-1)^{j-1}\sin((j-\alpha_j)\pi/n)}{n\sin\alpha\pi}\left(b_n + 2\sum_{k=1}^{n-1} b_{n-k}\cos\frac{k(j-\alpha_j)\pi}{n}\right). \tag{26}$$

4.2. Problem 2. For this problem the order of the linear systems is $n = 2p$. Because $T_{2p}(\lambda) = 2T_p^2(\lambda) - 1$, it follows immediately from (22) that

$$\begin{aligned}
&\frac{(T_p(\lambda)-\sigma)^2 - (T_p(\mu)-\sigma)^2}{\lambda-\mu} = \frac{T_p^2(\lambda) - T_p^2(\mu)}{\lambda-\mu} - 2\sigma\frac{T_p(\lambda)-T_p(\mu)}{\lambda-\mu}\\
&= \sum_{k=0}^{2p-2} T_{2p-k-1}(\lambda)U_k(\mu) + \frac{1}{2}T_0(\lambda)U_{2p-1}(\mu)\\
&\quad - 4\sigma\sum_{k=0}^{p-2} T_{p-k-1}(\lambda)U_k(\mu) - 2\sigma T_0(\lambda)U_{p-1}(\mu).
\end{aligned}$$

This equality shows that

$$\begin{aligned}
\{\widehat{U}_0(\lambda), \widehat{U}_1(\lambda), \ldots, \widehat{U}_{2p-1}(\lambda)\} = \Big\{&U_0(\lambda), \ldots, U_{p-1}(\lambda),\\
&U_p(\lambda) - 4\sigma U_0(\lambda), \ldots, U_{2p-2}(\lambda) - 4\sigma U_{p-2}(\lambda), \frac{1}{2}U_{2p-1}(\lambda) - 2\sigma U_{p-1}(\lambda)\Big\}
\end{aligned}$$

is the J-match of $\{T_0(\lambda), T_1(\lambda), \ldots, T_{n-1}(\lambda)\}$ with the link $(T_p(\lambda)-\sigma)^2$ and

$$\begin{aligned}
\{\widehat{T}_0(\lambda), \widehat{T}_1(\lambda), \ldots, \widehat{T}_{2p-1}(\lambda)\} = \Big\{&\frac{1}{2}T_0(\lambda), T_1(\lambda), \ldots, T_{p-1}(\lambda),\\
&T_p(\lambda) - 2\sigma T_0(\lambda), T_{p+1}(\lambda) - 4\sigma T_1(\lambda), \ldots, T_{2p-1}(\lambda) - 4\sigma T_{p-1}(\lambda)\Big\}
\end{aligned}$$

is the J-match of $\{U_0(\lambda), U_1(\lambda), \dots, U_{n-1}(\lambda)\}$ with the link $(T_p(\lambda)-\sigma)^2$. It is readily checked that $r(\lambda) = 2^{-2p+2}(T_p(\lambda)-\sigma)^2$. Hence, for the linear systems $V_{c,t}\mathbf{x} = \mathbf{b}$,

$$v(\lambda) = 2^{-2p+2}(b_1\widehat{U}_{2p-1}(\lambda) + b_2\widehat{U}_{2p-2}(\lambda) + \cdots + b_{2p-1}\widehat{U}_1(\lambda) + b_{2p}\widehat{U}_0(\lambda)),$$

and for the linear systems $V_{c,u}\mathbf{x} = \mathbf{b}$,

$$v(\lambda) = 2^{-2p+2}(b_1\widehat{T}_{2p-1}(\lambda) + b_2\widehat{T}_{2p-2}(\lambda) + \cdots + b_{2p-1}\widehat{T}_1(\lambda) + b_{2p}\widehat{U}_0(\lambda)).$$

Now our duty is to compute $(v(\lambda)/r_j(\lambda))^{(k)}$ for $\lambda = t_j$ and $k = 0, 1$. Denote the polynomial $q_j(\lambda) = (\lambda - t_1)\cdots(\lambda - t_{j-1})(\lambda - t_{j+1})\cdots(\lambda - t_p)$. It is straightforward to show that $r_j(\lambda) = q_j^2(\lambda)$ and $q_j(\lambda)(\lambda - t_j) = 2^{-p+1}(T_p(\lambda) - \sigma)$. Therefore, $q_j(t_j) = 2^{-p+1}T_p'(t_j)$ and $q_j'(t_j) = 2^{-p}T_p''(t_j)$. Applying Theorem 4.1 we obtain the solution of $V_{c,t}\mathbf{x} = \mathbf{b}$ and $V_{c,u}\mathbf{x} = \mathbf{b}$,

$$x_{2j-1} = \frac{b'(t_j)}{(T_p'(t_j))^2} - \frac{T_p''(t_j)b(t_j)}{(T_p'(t_j))^3}, \tag{27}$$

$$x_{2j} = \frac{b(t_j)}{(T_p'(t_j))^2}, \qquad j = 1, 2, \dots, p, \tag{28}$$

where

$$b(\lambda) = \begin{cases} b_1\widehat{U}_{2p-1}(\lambda) + \cdots + b_{2p-1}\widehat{U}_1(\lambda) + b_{2p}\widehat{U}_0(\lambda) & \text{for } V_{c,t}\mathbf{x} = \mathbf{b}, \\ b_1\widehat{T}_{2p-1}(\lambda) + \cdots + b_{2p-1}\widehat{T}_1(\lambda) + b_{2p}\widehat{T}_0(\lambda) & \text{for } V_{c,u}\mathbf{x} = \mathbf{b}. \end{cases}$$

4.3. Problem 3. For Problem 3, $n = 3p$. Because $T_{3p}(\lambda) = 4T_p^3(\lambda) - 3T_p(\lambda)$, it follows from (22) that

$$\begin{aligned}
&\frac{(T_p(\lambda)-\sigma)^3 - (T_p(\mu)-\sigma)^3}{\lambda-\mu} = \frac{1}{4}\frac{T_{3p}(\lambda) - T_{3p}(\mu)}{\lambda-\mu} \\
&- \frac{3}{2}\sigma\frac{T_{2p}(\lambda) - T_{2p}(\mu)}{\lambda-\mu} + \left(3\sigma^2 + \frac{3}{4}\right)\frac{T_p(\lambda) - T_p(\mu)}{\lambda-\mu} \\
&= \frac{1}{2}\sum_{k=0}^{3p-2} T_{3p-k-1}(\lambda)U_k(\mu) + \frac{1}{4}T_0(\lambda)U_{3p-1}(\mu) - 3\sigma\sum_{k=0}^{2p-2} T_{2p-k-1}(\lambda)U_k(\mu) \\
&- \frac{3\sigma}{2}T_0(\lambda)U_{2p-1}(\mu) + \left(6\sigma^2 + \frac{3}{2}\right)\sum_{k=0}^{p-2} T_{p-k-1}U_k(\mu) + \left(3\sigma^2 + \frac{3}{4}\right)T_0(\lambda)U_{3p-1}(\mu).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\{\widetilde{U}_0(\lambda), \widetilde{U}_1(\lambda), \dots, \widetilde{U}_{3p-1}(\lambda)\} = \Big\{&\frac{1}{2}U_0(\lambda), \dots, \frac{1}{2}U_{p-1}(\lambda), \\
&\frac{1}{2}U_p(\lambda) - 3\sigma U_0(\lambda), \dots, \frac{1}{2}U_{2p-1}(\lambda) - 3\sigma U_{p-1}(\lambda), \frac{1}{2}U_{2p}(\lambda) - 3\sigma U_p(\lambda) \\
&+ \left(6\sigma^2 + \frac{3}{2}\right)U_0(\lambda), \dots, \frac{1}{2}U_{3p-2}(\lambda) - 3\sigma U_{2p-2}(\lambda) + \left(6\sigma^2 + \frac{3}{2}\right)U_{p-2}(\lambda), \\
&\frac{1}{4}U_{3p-1}(\lambda) - \frac{3\sigma}{2}U_{2p-1}(\lambda) + \left(3\sigma^2 + \frac{3}{4}\right)U_{p-1}(\lambda)\Big\}
\end{aligned}$$

is the J-match of $\{T_0(\lambda),\dots,T_{3p-1}(\lambda)\}$ with the link $(T_p(\lambda)-\sigma)^3$ and

$$\begin{aligned}\{\widetilde{T}_0(\lambda),\widetilde{T}_1(\lambda),\dots,\widetilde{T}_{3p-1}(\lambda)\} &= \Big\{\frac{1}{4}T_0(\lambda),\frac{1}{2}T_1(\lambda),\dots,\frac{1}{2}T_{p-1}(\lambda),\\ &\frac{1}{2}T_p(\lambda)-\frac{3\sigma}{2}T_0(\lambda),\frac{1}{2}T_p(\lambda)-3\sigma T_1(\lambda),\dots,\frac{1}{2}T_{2p-1}(\lambda)-3\sigma T_{p-1}(\lambda),\\ &\frac{1}{2}T_{2p-1}(\lambda)-3\sigma T_{p-1}(\lambda)+\left(3\sigma^2+\frac{3}{4}\right)T_0(\lambda),\frac{1}{2}T_{2p}(\lambda)-3\sigma T_p(\lambda)\\ &+\left(6\sigma^2+\frac{3}{2}\right)T_1(\lambda),\dots,\frac{1}{2}T_{3p-1}(\lambda)-3\sigma T_{2p-1}(\lambda)+\left(6\sigma^2+\frac{3}{2}\right)T_{p-1}(\lambda)\Big\}\end{aligned}$$

is the J-match of $\{U_0(\lambda),U_1(\lambda),\dots,U_{3p-1}(\lambda)\}$ with the link $(T_p(\lambda)-\sigma)^3$. For this problem, $r(\lambda)=2^{-3p+3}(T_p(\lambda)-\sigma)^3$, which implies that for linear systems $V_{c,t}\mathbf{x}=\mathbf{b}$

$$v(\lambda)=2^{-3p+3}(b_1\widetilde{U}_{3p-1}(\lambda)+\cdots+b_{3p-1}\widetilde{U}_1(\lambda)+b_{3p}\widetilde{U}_0(\lambda)),$$

and for linear systems $V_{c,u}\mathbf{x}=\mathbf{b}$,

$$v(\lambda)=2^{-3p+3}(b_1\widetilde{T}_{3p-1}(\lambda)+\cdots+b_{3p-1}\widetilde{T}_1(\lambda)+b_{3p}\widetilde{T}_0(\lambda)).$$

Since $r_j(\lambda)=q_j^3(\lambda)$, applying Theorem 4.1 yields the solution of the systems

$$\begin{aligned}x_{3j-2} &= \frac{1}{2}\left(\frac{v(\lambda)}{r_j(\lambda)}\right)''\Big|_{\lambda=t_j}\\ &= \frac{1}{2}\left(\frac{b''(t_j)}{(T_p'(t_j))^3}-\frac{3b'(t_j)T_p''(t_j)+T_p'''(t_j)b(t_j)}{(T_p'(t_j))^4}+\frac{3b(t_j)(T_p''(t_j))^2}{(T_p'(t_j))^5}\right),\\ x_{3j-1} &= \left(\frac{v(\lambda)}{r_j(\lambda)}\right)\Big|_{\lambda=t_j}=\frac{b'(t_j)}{(T_p'(t_j))^3}-\frac{3}{2}\frac{b(t_j)T_p''(t_j)}{(T_p'(t_j))^4}\\ x_{3j} &= \frac{1}{2}\frac{b(t_j)}{(T_p'(t_j))^3},\qquad j=1,2,\dots,p,\end{aligned}$$

where

$$b(\lambda)=\begin{cases} b_1\widetilde{U}_{3p-1}(\lambda)+\cdots+b_{3p-1}\widetilde{U}_1(\lambda)+b_{3p}\widetilde{U}_0(\lambda) & \text{for } V_{c,t}\mathbf{x}=\mathbf{b},\\ b_1\widetilde{T}_{3p-1}(\lambda)+\cdots+b_{3p-1}\widetilde{T}_1(\lambda)+b_{3p}\widetilde{T}_0(\lambda) & \text{for } V_{c,u}\mathbf{x}=\mathbf{b},\end{cases}$$

Hence, for Problems 1, 2, and 3 the solution of confluent Chebyshev–Vandermonde systems $V_{c,t}\mathbf{x}=\mathbf{b}$ or $V_{c,u}\mathbf{x}=\mathbf{b}$ for Chebyshev σ-points is reduced to FFT or sine or cosine transforms, which yields $O(n\log n)$ algorithms. A detail implementation for Problem 2 of the systems $V_{c,t}\mathbf{x}=\mathbf{b}$ by FFT will be given in the appendix.

5. Application of the generalized Hilbert matrix problem to the solution of Chebyshev–Vandermonde systems. In this section, we consider an application of the generalized Hilbert matrix problem to the solution of Chebyshev–Vandermonde systems of the first and the second kinds if the points $t_k=\cos\frac{(k-1+\alpha)\pi}{n}$, $k=1,\dots,n$, $0\le\alpha<1$. If $\alpha=1/2$, then t_k become zeros of $T_n(\lambda)$. The corresponding systems are solved by the method in section 4.1. We assume that $\alpha\ne 1/2$. If $\alpha=0$, t_k become extreme points of $T_n(\lambda)$. To this end, we need the following result on the inversion of a class of generalized Hilbert matrices [6]. An alternative

proof by using the J-match and the link of polynomials is given here for the readers' convenience.

COROLLARY 5.1. *Let $H(\mathbf{r},\mathbf{t})$ be a generalized Hilbert matrix with the (k,j) entry $(H(\mathbf{r},\mathbf{t}))_{kj} = 1/(r_k - t_j)$, $k,j = 1,2,\ldots,n$, where r_k, t_j are distinct points in the complex plane. Then $H(\mathbf{r},\mathbf{t})$ is nonsingular and*

$$H^{-1}(\mathbf{r},\mathbf{t}) = D_1 H(\mathbf{t},\mathbf{r}) D_2, \tag{29}$$

where

$$\begin{gathered} D_1 = \operatorname{diag}(s(t_1)/l'(t_1), s(t_2)/l'(t_2), \ldots, s(t_n)/l'(t_n)), \\ D_2 = \operatorname{diag}(l(r_1)/s'(r_1), l(r_2)/s'(r_2), \ldots, l(r_n)/s'(r_n)), \\ l(\lambda) = (\lambda - t_1)(\lambda - t_2)\cdots(\lambda - t_n), \quad s(\lambda) = (\lambda - r_1)(\lambda - r_2)\cdots(\lambda - r_n). \end{gathered}$$

Proof. Denote $l_i(\lambda) = l(\lambda)/(\lambda - t_i)$. Then $\{l_i(\lambda)\}_{i=1}^n$ is a basis of $\mathbb{C}_{n-1}[\lambda]$. For the polynomial $s(\lambda)$, applying Corollary 3.1 shows that there exists a unique basis $p_i(\lambda)$, $i = 1,2,\ldots,n$, of $\mathbb{C}_{n-1}[\lambda]$ such that

$$\sum_{k=1}^{n} l_k(\lambda) p_{n-k+1}(\mu) = \frac{s(\lambda) - s(\mu)}{\lambda - \mu}. \tag{30}$$

Putting $\lambda = t_i$ in (30) shows that $p_{n-i+1}(\mu) = (s(\mu) - s(t_i))/((\mu - t_i) l_i(t_i))$. Choosing $\lambda = r_i$ and $\mu = r_j$ in (30) yields that

$$\sum_{k=1}^{n} l_k(r_i) \frac{s(t_k)}{(t_k - r_j) l_k(t_k)} = \begin{cases} s'(r_i), & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases} \tag{31}$$

On the other hand,

$$H(\mathbf{r},\mathbf{t}) = \operatorname{diag}(1/l(r_1), \ldots, 1/l(r_n)) \begin{pmatrix} l_1(r_1) & l_2(r_1) & \cdots & l_n(r_1) \\ l_1(r_2) & l_2(r_2) & \cdots & l_n(r_2) \\ \cdots & \cdots & \cdots & \cdots \\ l_1(r_n) & l_2(r_n) & \cdots & l_n(r_n) \end{pmatrix},$$

which, together with (31), yields the desired result. ∎

Now consider Chebyshev–Vandermonde systems $V\mathbf{x} = \mathbf{b}$ with $t_k = \cos\frac{(k-1)\pi + \alpha}{n}$, where V is the Chebyshev–Vandermonde matrix of the first or second kind. Let r_k be zeros of $T_n(\lambda) - \nu$, where the constant $\nu = \cos\beta\pi$ is chosen such that $|\nu| < 1$ and $T_n(t_i) \neq \nu$. Denote

$$\widetilde{V} = \begin{pmatrix} q_n(r_1) & q_{n-1}(r_1) & \cdots & q_1(r_1) \\ q_n(r_2) & q_{n-1}(r_2) & \cdots & q_1(r_2) \\ \cdots & \cdots & \cdots & \cdots \\ q_n(r_n) & q_{n-1}(r_n) & \cdots & q_1(r_n) \end{pmatrix},$$

where $\{q_1(\lambda), q_2(\lambda), \ldots, q_n(\lambda)\}$ is the J-match of the systems with the link $T_n(\lambda)$ given by (22). Applying Lemma 3.1 in [25] shows that $\widetilde{V}$ is nonsingular. Therefore, $\mathbf{x}$ is the solution of $V\mathbf{x} = \mathbf{b}$ if and only if $\mathbf{x}$ is the solution of the following systems:

$$\widetilde{V}V\mathbf{x} = \widetilde{V}\mathbf{b}. \tag{32}$$

Using (22) shows that $\widetilde{V}V = \mathrm{diag}(\nu - T_n(t_1), \ldots, \nu - T_n(t_n))H(\mathbf{r}, \mathbf{t})$. Furthermore, applying Corollary 5.1 and (20) with a simple computation yields the solution of (32),

$$\mathbf{x} = -2^{1-n}\widetilde{D}_1\widetilde{V}(\mathbf{p})V(\mathbf{q})D_2\widetilde{V}\mathbf{b}, \tag{33}$$

where $\widetilde{D}_1 = \mathrm{diag}(\frac{1}{l'(t_1)(T_n(t_1)-\cos\beta\pi)}, \ldots, \frac{1}{l'(t_n)(T_n(t_n)-\cos\beta\pi)})$ and D_2 is the same as in Corollary 5.1. The products of matrices involved in (33) with a vector can be computed by FFT or sine or cosine transforms. An elementary calculation shows that $s'(r_k) = 2^{1-n}(-1)^{k-1}\sin\beta\pi/\sin((k-\beta_k)\pi/n)$, where $\beta_k = 1-\beta$ if k is odd, $\beta_k = \beta$ if k is even. For $t_k = \cos\frac{(k-1+\alpha)\pi}{n}$, values of $l'(t_k)$ and $l(r_k)$ are considerably small even for a moderate n. Directly computing them leads easily to underflow. To obtain an accurate solution we compute diagonal matrices $2^{1-n}\widetilde{D}_1$ and D_2 globally instead of computing $l'(t_k)$ and $l(r_k)$. The approach is similar to the computation of $s(t_i)/l'(t_i)$ and $l(r_i)/s'(r_i)$ for generalized Hilbert systems in section 6.3. This is fulfilled by $O(n)$ operations. We delete further details. Therefore, the overall arithmetic operations for solution (33) are $O(n\log n)$.

6. Applications. In this section we consider applications of the results from the previous sections to some related problems.

6.1. Solution of confluent Chebyshev–Vandermonde systems for near Chebyshev σ-points. Distinct points $\{t_1, t_2, \ldots, t_p\}$ are near Chebyshev σ-points if there are only m ($m \ll p$) points t_i which are not zeros of $T_p(\lambda) - \sigma$. Confluent Chebyshev–Vandermonde systems with these kind of points often occur in practice, for example, $m = 1, 2$. For Problems 1, 2, and 3 of the systems for near Chebyshev σ-points, J-matches and links are chosen the same as used in section 4. Let $\{r_1, r_2, \ldots, r_p\}$ are all zeros of $T_p(\lambda) - \sigma$ and assume

$$\{t_1, t_2, \ldots, t_p\} \setminus \{r_1, r_2, \ldots, r_p\} = \{t_1, \ldots, t_m\},$$
$$\{r_1, r_2, \ldots, r_p\} \setminus \{t_1, t_2, \ldots, t_p\} = \{r_1, \ldots, r_m\}.$$

For Problem k ($k = 1, 2, 3$) it follows from the discussion in section 4 that

$$v(\lambda) = 2^{-n+k}\mathrm{quot}((\widetilde{q}(\lambda))^k b(\lambda), (\widetilde{p}(\lambda))^k),$$

where $\widetilde{q}(\lambda) = (\lambda - t_1)\cdots(\lambda - t_m)$ and $\widetilde{p}(\lambda) = (\lambda - r_1)\cdots(\lambda - r_m)$. The polynomial $b(\lambda)$ is easily represented by

$$b(\lambda) = \begin{cases} d_{n-1}U_{n-1}(\lambda) + \cdots + d_0U_0(\lambda) & \text{for } V_{c,t}\mathbf{x} = \mathbf{b}, \\ \widetilde{d}_{n-1}T_{n-1}(\lambda) + \cdots + \widetilde{d}_0T_0(\lambda) & \text{for } V_{c,u}\mathbf{x} = \mathbf{b}, \end{cases} \tag{34}$$

in $O(n)$ arithmetic operations. We consider linear systems $V_{c,t}\mathbf{x} = \mathbf{b}$ only. In order to represent $v(\lambda)$ in terms of $\{U_i(\lambda)\}_{i=0}^{n-1}$ we need the following lemma.

LEMMA 6.1. *Let $p(\lambda), q(\lambda)$, $r(\lambda)$ be polynomials, $p(\lambda) = p_1(\lambda)p_2(\lambda)$, and $q(\lambda) = q_1(\lambda)q_2(\lambda)$ with $\deg(p_2(\lambda)) \le \deg(q_2(\lambda))$, $s_1(\lambda) = \mathrm{quot}(p_1(\lambda)r(\lambda), q_1(\lambda))$, and $s(\lambda) = \mathrm{quot}(p_2(\lambda)s_1(\lambda), q_2(\lambda))$. Then $s(\lambda) = \mathrm{quot}(p(\lambda)r(\lambda), q(\lambda))$.*

Proof. Assume $p_1(\lambda)r(\lambda) = s_1(\lambda)q_1(\lambda)+r_1(\lambda)$ and $p_2(\lambda)s_1(\lambda) = s(\lambda)q_2(\lambda)+r_2(\lambda)$, where $\deg(r_1(\lambda)) < \deg(q_1(\lambda))$ and $\deg(r_2(\lambda)) < \deg(q_2(\lambda))$. Then

$$\begin{aligned} p(\lambda)r(\lambda) &= p_2(\lambda)p_1(\lambda)r(\lambda) = p_2(\lambda)s_1(\lambda)q_1(\lambda) + p_2(\lambda)r_1(\lambda) \\ &= s(\lambda)q(\lambda) + r_2(\lambda)q_1(\lambda) + p_2(\lambda)r_1(\lambda). \end{aligned}$$

Under the assumption it is readily seen that $\deg(r_2(\lambda)q_1(\lambda)+p_2(\lambda)r_1(\lambda)) < \deg(q(\lambda))$. Therefore, $s(\lambda) = \mathrm{quot}(p(\lambda)r(\lambda), q(\lambda))$. □

Now we consider the computation of $\mathrm{quot}(b(\lambda)(\lambda-a), \lambda-d)$, where a and d are constants. Denoting $c_i = d_i a$, $i = 0, 1, \ldots, n-1$, we have

$$\begin{aligned}(\lambda-a)b(\lambda) &= d_{n-1}\lambda U_{n-1}(\lambda) + d_{n-2}\lambda U_{n-2}(\lambda) + \cdots + d_0\lambda U_0(\lambda)\\ &\quad - c_{n-1}U_{n-1}(\lambda) - c_{n-1}U_{n-2}(\lambda) - \cdots - c_0U_0(\lambda)\\ &= d_{n-1}(\lambda-d)U_{n-1}(\lambda) + d_{n-2}\lambda U_{n-2}(\lambda) + \cdots + d_0\lambda U_0(\lambda)\\ &\quad + (d_{n-1}d - c_{n-1})U_{n-1}(\lambda) - c_{n-2}(\lambda)U_{n-2}(\lambda) - \cdots - c_0U_0(\lambda)\\ &= d_{n-1}(\lambda-d)U_{n-1}(\lambda) + (d_{n-2} + 2(d_{n-1}d - c_{n-1}))\lambda U_{n-2}(\lambda)\\ &\quad + d_{n-3}\lambda U_{n-3}(\lambda) + \cdots + d_0\lambda U_0(\lambda) - c_{n-2}U_{n-2}(\lambda)\\ &\quad - (c_{n-3} + d_{n-1}d - c_{n-1})U_{n-3}(\lambda) - c_{n-4}U_{n-4}(\lambda) - \cdots - c_0U_0(\lambda),\end{aligned}$$

because $U_i(\lambda) = 2\lambda U_{i-1}(\lambda) - U_{i-2}(\lambda)$ for $i \geq 2$. Therefore, $\mathrm{quot}((\lambda-a)b(\lambda), x-d)$ is computed as follows.

ALGORITHM 2. Let $b(\lambda) = d_{n-1}U_{n-1}(\lambda) + \cdots + d_0U_0(\lambda)$. Algorithm 2 represents $\mathrm{quot}((\lambda-a)b(\lambda), x-d)$ in term of the basis $\{U_i(\lambda)\}_{i=0}^{n-1}$, where a and d are constants.

$c_i = d_i a$, $i = 1, \ldots, n-1$.
For $i = n-2 : -1 : 0$
  $d_i = d_i + 2(d_{i+1}d - c_{i+1})$
  if $i > 1$ then
    $c_{i-1} = c_{i-1} + d_{i+1}d - c_{i+1}$
  endif
endfor i.

We then have $\mathrm{quot}((\lambda-a)b(\lambda), x-d) = \sum_{k=0}^{n-1} d_k U_k(\lambda)$. Algorithm 2 needs $O(n)$ arithmetic operations. Using Algorithm 2 km times we represent $v(\lambda)$ in terms of the basis $\{U_i(\lambda)\}_{i=0}^{n-1}$ for problem k of the systems $V_{c,t}\mathbf{x} = \mathbf{b}$. Similarly, the polynomial $v(\lambda)$ for Problem k of $V_{c,u}\mathbf{x} = \mathbf{b}$ can be represented in terms of $\{T_i(\lambda)\}_{i=0}^{n-1}$ in $O(mn)$ arithmetic operations. Using the same notation in section 4 we have $r_i(\lambda) = (q_i(\lambda))^k$ for problem k and

$$q_i(\lambda)(\lambda - t_i) = 2^{-p+1}\frac{(\lambda-t_1)\cdots(\lambda-t_m)}{(\lambda-r_1)\cdots(\lambda-r_m)}(T_p(\lambda) - \sigma).$$

By using Theorem 4.1, the same approach in section 4 is applicable to the problems in this subsection. If $m \ll n$, it is readily seen that all components of the solution determined by the zeros of $T_p(\lambda) - \sigma$ can be computed in $O(mn + n\log n)$ arithmetic operations while the rest are computed in $O(nm)$ operations. We omit further details. The total operations are $O(nm + n\log n)$. Clearly, Algorithm 2 is applicable to any points for confluent Chebyshev–Vandermonde systems. This approach yields a new $O(n^2)$ algorithm for any choice of t_k.

6.2. Dual confluent Chebyshev–Vandermonde systems. The results in sections 4 and 5 are applicable to the corresponding dual Chebyshev–Vandermonde systems, which are equivalent to Hermite interpolating problems in terms of basis $\{T_k(\lambda)\}_{k=0}^{n-1}$ or $\{U_k(\lambda)\}_{k=0}^{n-1}$.

Consider dual Chebyshev–Vandermonde systems $V^T\mathbf{x} = \mathbf{b}$ with $t_k = \cos\frac{(k-1)\pi+\alpha}{n}$, where V is the Chebyshev–Vandermonde matrix of the first or the second kind. It follows from (33) that $V^{-1} = -2^{1-n}\widetilde{D}_1\widetilde{V}(\mathbf{p})V(\mathbf{q})D_2\widetilde{V}$. Hence, the solution of $V^T\mathbf{x} = \mathbf{b}$

is given by

$$\mathbf{x} = -2^{1-n}(\widetilde{V})^T D_2 (V(\mathbf{q}))^T (\widetilde{V}(\mathbf{p}))^T \widetilde{D}_1 \mathbf{b},$$

which can also be computed in $O(n \log n)$ arithmetic operations by FFT or sine or cosine transforms.

For Problem 1 of dual Chebyshev–Vandermonde systems for Chebyshev σ-points, it straightforwardly follows from (25) that

$$V_{c,t}^{-1} = \frac{1}{n \sin \alpha\pi} \mathrm{diag}(1, -1, \ldots, (-1)^{n-1})$$
$$\begin{pmatrix} \sin\alpha_1\pi & 2\sin\frac{(n-1)(1-\alpha_1)\pi}{n} & 2\sin\frac{(n-2)(1-\alpha_1)\pi}{n} & \ldots & 2\sin\frac{(1-\alpha_1)\pi}{n} \\ -\sin\alpha_2\pi & 2\sin\frac{(n-1)(2-\alpha_2)\pi}{n} & 2\sin\frac{(n-2)(2-\alpha_2)\pi}{n} & \ldots & 2\sin\frac{(2-\alpha_2)\pi}{n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ (-1)^{n-1}\sin\alpha_n\pi & 2\sin\frac{(n-1)(n-\alpha_n)\pi}{n} & 2\sin\frac{(n-2)(n-\alpha_n)\pi}{n} & \ldots & 2\sin\frac{(n-\alpha_n)\pi}{2n} \end{pmatrix}.$$

Therefore, the solution of the dual Vandermonde–Chebyshev linear systems $V_{c,t}^T \mathbf{x} = \mathbf{b}$ is given by

$$x_1 = \frac{1}{n}\sum_{k=1}^{n} b_k, \quad x_j = \frac{2}{n \sin \alpha\pi} \sum_{k=1}^{n} (-1)^{k-1} b_k \sin \frac{(n-j+1)(k-\alpha_k)\pi}{n}, \quad j > 1.$$

Similarly, it follows from (26) that the solution for $V_{c,u}^T \mathbf{x} = \mathbf{b}$ for Problem 1 of Chebyshev σ-points is given by

$$x_j = \frac{2}{n}\sum_{k=1}^{n} d_k \cos \frac{(n-j)(k-\alpha_k)\pi}{n}, \quad j < n, \qquad x_n = \frac{1}{n}\sum_{k=1}^{n} d_k,$$

where $(d_1, d_2, \ldots, d_n)^T = \frac{1}{\sin \alpha\pi} D\mathbf{b}$ and

$$D = \mathrm{diag}(\sin(1-\alpha_1)\pi/n, -\sin(2-\alpha_2)\pi/n, \ldots, (-1)^{n-1}\sin(n-\alpha_n)\pi/n).$$

For Problems 2 and 3 of dual confluent Chebyshev–Vandermonde systems for Chebyshev σ-points, we can solve them in similar ways. Based on the solution of confluent Chebyshev–Vandermonde systems in section 4, it is not hard to obtain the inversion of $V_{c,t}$ and $V_{u,t}$ by choosing the right-hand side $\mathbf{e}_i = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{in})^T$ for $i = 1, \ldots, n$, where δ_{ij} is the Kronecker δ-function. Another way to obtain the inversion of $V_{c,t}$ and $V_{u,t}$ is to use Corollary 3.4 in [25]. Then the solution of $V_{c,t}^T \mathbf{x} = \mathbf{b}$ and $V_{c,u}^T \mathbf{x} = \mathbf{b}$ for Problems 2 and 3 are given by $\mathbf{x} = ((V_{c,t})^{-1})^T \mathbf{b}$ and $\mathbf{x} = ((V_{c,u})^{-1})^T \mathbf{b}$, respectively. Both can be computed by FFT or sine or cosine transforms that lead to $O(n \log n)$ algorithms. We omit further details.

6.3. Solution of a class of generalized Hilbert linear systems. In this subsection the result on the generalized Hilbert matrix problem in section 2 is applied to a class of generalized Hilbert systems

$$H(\mathbf{r}, \mathbf{t})\mathbf{x} = \mathbf{b}, \tag{35}$$

where $(H(\mathbf{r}, \mathbf{t}))_{i,j} = 1/(r_i - t_j)$ with $r_i \neq t_j$, $i, j = 1, 2, \ldots, n$. It is shown that there is an $O(n \log^2 n)$ algorithm for the systems in general [22]. A generalization of the systems is found in [14]. We consider how to obtain an $O(n \log n)$ algorithm for the

case where $t_i = f(i)$ and $r_j = g(j)$, and where function pairs (f, g), (f, f), and (g, g) satisfy TH-relations. Applying Corollary 5.1 we have

$$\mathbf{x} = D_1 H(\mathbf{t}, \mathbf{r}) D_2 \mathbf{b}.$$

Under the assumption that (f, g), (f, f), and (g, g) satisfy TH-relations, it is readily shown that for $x \neq y$ there exist functions $\widetilde{d}_1, \widetilde{d}_2, \widetilde{q}_1, \widetilde{q}_2, \widehat{d}_1, \widehat{d}_2, \widehat{q}_1, \widehat{q}_2$ such that

$$\frac{f(x) - g(y)}{f(x) - f(y)} = \widetilde{d}_1(x)\widetilde{d}_2(y)\widetilde{q}_1(x+y)\widetilde{q}_2(x-y),$$
$$\frac{g(x) - f(y)}{g(x) - g(y)} = \widehat{d}_1(x)\widehat{d}_2(y)\widehat{q}_1(x+y)\widehat{q}_2(x-y).$$

Because $r_i \neq r_j$, $r_i \neq t_j$, and $t_i \neq t_j$ for $i \neq j$ we have $\widetilde{d}_1(i) \neq 0$, $\widetilde{d}_2(i) \neq 0$, $\widehat{d}_1(i) \neq 0$, $\widehat{d}_2(i) \neq 0$, $i = 1, 2, \ldots, n$ and choosing $x = i + 1, y = i - 1$ shows that for $1 < i < n$,

$$\widetilde{q}_1(2i) \neq 0, \quad \widehat{q}_1(2i) \neq 0. \tag{36}$$

For convenience, we assume also that (36) holds for $i = 1, n$. Hence,

$$\begin{aligned}
\frac{s(t_i)}{l'(t_i)} &= \frac{t_i - r_i}{\widetilde{d}_2(i)\widetilde{q}_1(2i)} (\widetilde{d}_1(i))^{n-1} \prod_{j=1}^{n} \widetilde{d}_2(j)\widetilde{q}_1(i+j) \prod_{j \neq i} \widetilde{q}_2(i-j) \\
&= \frac{t_i - r_i}{\widetilde{d}_2(i)\widetilde{q}_1(2i)} (\widetilde{d}_1(i))^{n-1} \frac{\widetilde{Q}(i+n)}{\widetilde{Q}(i)} \widetilde{P}(i-1)\widetilde{R}(n-i) \prod_{j=1}^{n} \widetilde{d}_2(j), \\
\frac{l(r_i)}{s'(r_i)} &= \frac{r_i - t_i}{\widehat{d}_1(i)\widehat{q}_1(2i)} (\widehat{d}_2(i))^{n-1} \prod_{j=1}^{n} \widehat{d}_1(j)\widehat{q}_1(i+j) \prod_{j \neq i} \widehat{q}_2(i-j) \\
&= \frac{r_i - t_i}{\widehat{d}_1(i)\widehat{q}_1(2i)} (\widehat{d}_2(i))^{n-1} \frac{\widehat{Q}(i+n)}{\widehat{Q}(i)} \widehat{P}(i-1)\widehat{R}(n-i) \prod_{j=1}^{n} \widehat{d}_1(j),
\end{aligned}$$

where

$$\widetilde{Q}(k) = \prod_{i=2}^{k} \widetilde{q}_1(i), \quad \widetilde{Q}(1) = 1, \quad \widetilde{P}(k) = \prod_{i=1}^{k} \widetilde{q}_2(i), \quad \widetilde{R}(k) = \prod_{i=1}^{k} \widetilde{q}_2(-i),$$
$$\widehat{Q}(k) = \prod_{i=2}^{k} \widehat{q}_1(i), \quad \widehat{Q}(1) = 1, \quad \widehat{P}(k) = \prod_{i=1}^{k} \widehat{q}_2(i), \quad \widehat{R}(k) = \prod_{i=1}^{k} \widehat{q}_2(-i).$$

For a complex number a it is well known that the power a^n can be computed by $O(\log n)$ arithmetic operations. Hence, diagonal matrices D_1 and D_2 can be computed in $O(n \log n)$ operations. Applying the result in section 2 for the generalized Hilbert matrix problem yields an $O(n \log n)$ algorithm for this class of generalized Hilbert systems.

Example 1. For the equidistant points $r_i = i/n$ and $t_i = (2i-1)/2n$, $i = 1, \ldots, n$, it is readily seen that $\widetilde{d}_1(x) = \widetilde{d}_2(x) = \widetilde{q}_1(x) = \widehat{d}_1(x) = \widehat{d}_2(x) = \widehat{q}_1(x) = 1$ and

$$\widetilde{q}_2(x) = 1 - \frac{1}{2x}, \qquad \widehat{q}_2(x) = 1 + \frac{1}{2x}.$$

Example 2. For the clustered points $r_i = (i/n)^2$ and $t_i = ((2i-1)/2n)^2$, $i = 1, \ldots, n$, we have $\widetilde{d}_1(x) = \widetilde{d}_2(x) = \widehat{d}_1(x) = \widehat{d}_2(x) = 1$ and

$$\widetilde{q}_1(x) = 1 + \frac{1}{2(x-1)}, \quad \widetilde{q}_2(x) = 1 - \frac{1}{2x}, \quad \widehat{q}_1(x) = 1 - \frac{1}{2x}, \quad \widehat{q}_2(x) = 1 + \frac{1}{2x}.$$

7. Numerical examples. In this section we present some numerical examples. For linear systems $A\mathbf{x} = \mathbf{b}$ we determined the relative residuals

$$\text{RES} = \frac{\|\mathbf{b} - A\mathbf{x}^{(dp)}\|_\infty}{\|A\|_\infty \|\mathbf{x}^{(dp)}\|_\infty + \|\mathbf{b}\|_\infty}, \qquad \text{QRES} = \frac{\|\mathbf{b} - A\mathbf{x}^{(dp)}\|_\infty}{\|b\|_\infty},$$

where $\mathbf{x}^{(dp)}$ stands for the solution computed in double precision arithmetic. RES is the ∞-norm backward error (see [19, Theorem 7.1]). The computations were performed on a SUN10 using IEEE-standard double precision arithmetic for which unit roundoff is $u_{dp} = 2^{-53} \approx 1.11 \times 10^{-16}$. The function randn$(n, 1)$ is used to generate n random numbers which are normally distributed with mean 0 and variance 1. All numerical examples are performed with the right-hand side $\mathbf{b} = ((-1).\hat{}(0 : n-1))^T$ and $\mathbf{b} = 10\text{randn}(n, 1)$. For $\mathbf{b} = 10\text{randn}(n, 1)$, we choose the maximum RES and corresponding QRES of 10 performances.

For the first numerical experiment we consider Problem 2 of $V_{c,t}\mathbf{x} = \mathbf{b}$ and $V_{c,u}\mathbf{x} = \mathbf{b}$ for Chebyshev σ-points by using the method in section 4.2 and the further discussion in the appendix. Tables 1–4 show the corresponding residuals. If α is near to 1 or 0, some roots of $T_p(\lambda) - \sigma$ are very close, where $\sigma = \cos \alpha\pi$. The systems are near singular. Our methods, however, still achieve accurate solutions, for example, $\alpha = 0.999$.

Since the methods for the generalized Hilbert matrix problem in sections 2 and 3 are applied to Chebyshev–Vandermonde systems and generalized Hilbert systems, for brief display, we give numerical examples for Chebyshev–Vandermonde systems for $t_k = \cos \frac{(k-1+\alpha)\pi}{n}$ by method in section 5 and generalized Hilbert linear systems $H(\mathbf{r}, \mathbf{t})\mathbf{x} = \mathbf{b}$ for the equidistant points $r_k = k/n$, $t_k = (2k-1)/(2n)$ and the clustered points $r_k = (k/n)^2$, $t_k = ((2k-1)/(2n))^2$ by the method in section 6.3. Tables 5–10 give the numerical errors. The numerical results show that the stability of the methods depends on point distributions. RES gets several orders of magnitude larger than u_{dp} for some points in Tables 5, 6, 7, and 10, indicating some instability.

8. Conclusions. The approach for Chebyshev–Vandermonde systems in section 4.1 is easily extended to the systems for the case where t_k are zeros of

$$p(\lambda) = a_0 T_n(\lambda) + a_1 T_{n-i_1}(\lambda) + \cdots + a_m T_{n-i_m}(\lambda),$$

where $a_0 \neq 0$, $i_1, \ldots, i_m$ are distinct positive integers, provided $t_1, t_2, \ldots, t_n$ are different. Denote $i_0 = 0$. We have

$$\frac{p(\lambda) - p(\mu)}{\lambda - \mu} = \sum_{k=0}^{m} a_k \frac{T_{n-i_k}(\lambda) - T_{n-i_k}(\mu)}{\lambda - \mu}. \tag{37}$$

By using (22) and (37) it is not hard to derive J-matches for $\{T_0(\lambda), \ldots, T_{n-1}(\lambda)\}$ and $\{U_0(\lambda), \ldots, U_{n-1}(\lambda)\}$ with the link $p(\lambda)$, respectively. Therefore, (24) gives the solution of Chebyshev–Vandermonde systems if t_k are zeros of $p(\lambda)$. In practice, $b(t_k)$, $k = 1, \ldots, n$ can often be computed by FFT or sine or cosine transforms, for example, Chebyshev–Gauss–Radau points $t_k = \cos \frac{2(k-1)\pi}{2n-1}$, i.e., zeros of $T_n(\lambda) + T_{n-1}(\lambda)$.

Based on the results in [25] the approach in section 4.1 is also applicable to other Vandermonde-like systems. In practice, the important fact for Vandermonde-like matrices and confluent Vandermonde-like matrices is that polynomials $p_1(\lambda), \ldots, p_n(\lambda)$ satisfy a k-term recurrence relation

$$p_1(\lambda) = 1, \quad p_i(\lambda) = 0, \quad i \leq 0, \tag{38}$$

TABLE 1
Numerical errors for Problem 2 of $V_{c,t}\mathbf{x} = \mathbf{b}$ *with* $\mathbf{b} = ((-1).\hat{}(0:n-1))^T$.

		$\alpha = 1/4$		$\alpha = 1/2$		$\alpha = 0.999$	
k	$n = 2^k$	RES	QRES	RES	QRES	RES	QRES
6	64	1.2e-19	1.0e-15	3.2e-19	8.9e-16	2.1e-19	7.2e-08
8	256	8.5e-21	1.3e-15	4.7e-20	2.7e-15	2.6e-20	1.9e-07
10	1024	1.8e-21	5.4e-15	1.4e-20	1.6e-14	5.2e-21	7.5e-07
12	4096	7.9e-22	4.4e-14	3.2e-21	6.7e-14	1.3e-21	3.5e-06
14	16384	1.9e-22	1.8e-13	5.1e-22	2.0e-13	2.6e-22	1.4e-05
16	65536	5.9e-23	1.1e-12	2.4e-22	1.8e-12	1.1e-22	1.1e-04

TABLE 2
Numerical errors for Problem 2 of $V_{c,t}\mathbf{x} = \mathbf{b}$ *with* $\mathbf{b} = 10\text{randn}(n, 1)$.

		$\alpha = 1/4$		$\alpha = 1/2$		$\alpha = 0.999$	
k	$n = 2^k$	RES	QRES	RES	QRES	RES	QRES
6	64	6.6e-19	1.7e-15	1.2e-18	7.6e-16	6.6e-19	1.7e-08
8	256	8.2e-20	2.4e-15	2.9e-19	1.4e-15	8.3e-20	3.3e-08
10	1024	2.8e-20	8.4e-15	1.3e-19	8.0e-15	3.9e-20	1.8e-07
12	4096	1.2e-20	3.4e-14	3.5e-20	1.8e-14	1.1e-20	5.1e-07
14	16384	3.6e-21	1.1e-13	1.2e-20	6.0e-14	3.6e-21	1.8e-06
16	65536	3.2e-21	8.1e-13	1.4e-20	6.5e-13	2.9e-21	1.3e-05

TABLE 3
Numerical errors for Problem 2 of $V_{c,u}\mathbf{x} = \mathbf{b}$ *with* $\mathbf{b} = ((-1).\hat{}(0:n-1))^T$.

		$\alpha = 1/4$		$\alpha = 1/2$		$\alpha = 0.999$	
k	$n = 2^k$	RES	QRES	RES	QRES	RES	QRES
6	64	1.2e-19	1.8e-15	1.3e-19	6.7e-16	8.5e-19	4.2e-08
8	256	1.5e-20	3.5e-15	2.4e-20	2.0e-15	1.0e-19	8.3e-08
10	1024	3.9e-21	1.4e-14	8.3e-21	1.1e-14	1.3e-20	1.7e-07
12	4096	8.3e-22	4.9e-14	2.9e-21	6.2e-14	4.1e-21	8.4e-07
14	16384	2.0e-22	1.9e-13	6.1e-22	2.1e-13	7.8e-22	2.5e-06
16	65536	1.1e-22	1.7e-12	2.4e-22	1.3e-12	2.1e-22	1.1e-05

TABLE 4
Numerical errors for Problem 2 of $V_{c,u}\mathbf{x} = \mathbf{b}$ *with* $\mathbf{b} = 10\text{randn}(n, 1)$.

		$\alpha = 1/4$		$\alpha = 1/2$		$\alpha = 0.999$	
k	$n = 2^k$	RES	QRES	RES	QRES	RES	QRES
6	64	4.4e-20	1.7e-15	7.6e-20	8.0e-16	1.7e-19	2.1e-08
8	256	1.5e-21	2.4e-15	3.2e-21	1.7e-15	3.9e-21	3.6e-08
10	1024	1.4e-22	9.7e-15	5.5e-22	6.8e-15	6.1e-22	1.7e-07
12	4096	1.7e-23	3.1e-14	4.7e-23	2.2e-14	4.9e-23	4.6e-07
14	16384	1.7e-24	8.8e-14	5.2e-24	6.9e-14	5.0e-24	1.7e-06
16	65536	4.5e-25	8.5e-13	1.5e-24	8.4e-13	1.2e-24	1.4e-05

$$p_{i+1}(\lambda) = \alpha_i(\lambda - \beta_i)p_i(\lambda) + \sum_{j=2}^{k-1} \gamma_{ij} p_{i-j+1}(\lambda), \qquad i = 1, \ldots, n, \tag{39}$$

where $\alpha_i \neq 0$, and t_k are often chosen zeros of $p_{n+1}(\lambda)$. Define a k-term recurrence relation by

$$q_1(\lambda) = 1, \quad q_i(\lambda) = 0, \quad i \leq 0, \tag{40}$$

TABLE 5
Numerical errors of Chebyshev–Vandermonde systems $V_{c,t}\mathbf{x} = \mathbf{b}$ *with* $\mathbf{b} = ((-1).\hat{}(0:n-1))^T$.

		$\alpha = 1/4$		$\alpha = 5/8$		$\alpha = 7/8$	
k	$n = 2^k$	RES	QRES	RES	QRES	RES	QRES
6	64	2.9e-15	1.9e-13	9.9e-16	6.4e-14	2.4e-16	1.5e-14
8	256	2.4e-14	6.2e-12	1.1e-15	2.7e-13	1.5e-16	3.8e-14
10	1024	3.0e-13	3.1e-10	2.9e-15	2.9e-12	2.9e-16	3.0e-13
12	4096	2.3e-12	9.4e-09	7.9e-15	3.2e-11	2.1e-16	8.5e-13
14	16384	2.4e-11	3.9e-07	2.7e-14	4.5e-10	2.1e-16	3.4e-12
16	65536	1.8e-10	1.2e-05	9.8e-14	6.4e-09	5.9e-16	3.8e-11

TABLE 6
Numerical errors of Chebyshev–Vandermonde systems $V_{c,t}\mathbf{x} = \mathbf{b}$ *with* $\mathbf{b} = 10\mathrm{randn}(n, 1)$.

		$\alpha = 1/4$		$\alpha = 5/8$		$\alpha = 7/8$	
k	$n = 2^k$	RES	QRES	RES	QRES	RES	QRES
6	64	6.6e-16	4.3e-14	3.7e-16	2.4e-14	2.0e-15	1.3e-13
8	256	1.1e-15	2.8e-13	4.8e-16	1.2e-13	1.3e-14	3.2e-12
10	1024	5.3e-15	5.4e-12	3.1e-15	3.2e-12	9.0e-14	9.3e-11
12	4096	2.9e-14	1.2e-10	9.5e-15	3.9e-11	4.0e-13	1.6e-09
14	16384	1.5e-13	2.5e-09	2.6e-14	4.3e-10	5.2e-12	8.5e-08
16	65536	7.9e-13	5.2e-08	8.3e-14	5.4e-09	3.7e-11	2.4e-06

TABLE 7
Numerical errors of Chebyshev–Vandermonde systems $V_{c,u}\mathbf{x} = \mathbf{b}$ *with* $\mathbf{b} = ((-1).\hat{}(0:n-1))^T$.

		$\alpha = 1/4$		$\alpha = 5/8$		$\alpha = 7/8$	
k	$n = 2^k$	RES	QRES	RES	QRES	RES	QRES
6	64	7.8e-17	1.4e-14	1.5e-16	3.4e-14	3.4e-16	5.8e-14
8	256	9.9e-16	8.9e-13	3.0e-17	3.2e-14	1.9e-16	1.5e-13
10	1024	1.5e-15	6.4e-12	7.6e-17	3.9e-13	1.5e-16	5.6e-13
12	4096	1.8e-15	3.5e-11	4.1e-17	9.7e-13	1.0e-16	1.8e-12
14	16384	3.7e-14	3.2e-09	2.7e-17	3.0e-12	1.4e-17	1.2e-12
16	65536	7.1e-13	4.3e-08	9.3e-17	4.5e-11	2.0e-16	7.3e-11

$$q_{i+1}(\lambda) = \alpha_{n-i+1}(\lambda - \beta_{n-i+1})q_i(\lambda) + \sum_{j=2}^{k-1} \gamma_{n-i+j,j} q_{i-j+1}(\lambda), \qquad i = 1, \ldots, n. \tag{41}$$

It is shown that $p_{n+1}(\lambda) = q_{n+1}(\lambda)$ and the basis $\{q_1(\lambda), \ldots, q_n(\lambda)\}$ is the J-match of $\{p_1(\lambda), \ldots, p_n(\lambda)\}$ with the link $p_{n+1}(\lambda)$ [25]. Let $b(\lambda) = b_1 q_n(\lambda) + \cdots + b_n q_1(\lambda)$. If the zeros of $p_{n+1}(\lambda)$ are distinct, it follows from (24) that the solution of Vandermonde-like systems $V(\mathbf{p})\mathbf{x} = \mathbf{b}$ is given by

$$x_k = b(t_k)/p'_{n+1}(t_k),$$

which is easily computed and often gives an accurate solution.

For Vandermonde-like systems and confluent Vandermonde-like systems, the stability of numerical methods is strongly influenced by points t_k. As we have seen, for proper points t_k we can obtain quite accurate results by using J-matches and links of polynomials. For an arbitrary choice of t_k a possible way to obtain an accurate solution is to use this approach, together with iterative refinement [18] to enhance the stability.

Appendix. Implementation of Problem 2. We now implement Problem 2 of $V_{c,t}\mathbf{x} = \mathbf{b}$ by using FFT efficiently. For this problem of Chebyshev σ-points, i.e.,

TABLE 8
Numerical errors of Chebyshev–Vandermonde systems $V_{c,u}\mathbf{x} = \mathbf{b}$ with $\mathbf{b} = 10\text{randn}(n, 1)$.

		$\alpha = 1/4$		$\alpha = 5/8$		$\alpha = 7/8$	
k	$n = 2^k$	RES	QRES	RES	QRES	RES	QRES
6	64	4.8e-17	8.9e-15	1.9e-17	4.2e-15	1.0e-16	1.7e-14
8	256	3.1e-16	2.8e-13	7.9e-18	8.5e-15	7.9e-17	6.4e-14
10	1024	1.2e-16	4.9e-13	8.0e-18	4.1e-14	6.9e-17	2.7e-13
12	4096	6.7e-17	1.3e-12	5.5e-18	1.3e-13	5.9e-17	1.0e-12
14	16384	6.6e-16	5.8e-11	8.7e-18	9.5e-13	1.1e-16	9.2e-12
16	65536	8.2e-16	3.2e-10	8.5e-18	4.2e-12	1.9e-16	6.9e-11

TABLE 9
Numerical errors of $H(\mathbf{r}, \mathbf{t})\mathbf{x} = \mathbf{b}$ for the equidistant points.

		$\mathbf{b} = ((-1).^\wedge(0 : n-1))^T$		$\mathbf{b} = 10\text{randn}(n, 1)$	
k	$n = 2^k$	RES	QRES	RES	QRES
6	64	2.2e-17	1.5e-14	1.0e-17	7.3e-15
8	256	1.0e-16	3.6e-13	5.6e-17	1.9e-13
10	1024	1.5e-16	2.5e-12	1.4e-16	2.4e-12
12	4096	1.8e-16	1.4e-11	1.9e-16	1.5e-11
14	16384	3.2e-15	1.1e-09	1.5e-15	5.5e-10
16	65536	9.4e-15	1.5e-08	4.2e-15	6.8e-09

TABLE 10
Numerical errors of $H(\mathbf{r}, \mathbf{t})\mathbf{x} = \mathbf{b}$ for the clustered points.

		$\mathbf{b} = ((-1).^\wedge(0 : n-1))^T$		$\mathbf{b} = 10\text{randn}(n, 1)$	
k	$n = 2^k$	RES	QRES	RES	QRES
6	64	7.2e-15	5.1e-12	4.8e-15	3.5e-12
8	256	3.1e-13	1.1e-09	1.6e-13	5.7e-10
10	1024	4.1e-12	7.0e-08	2.9e-12	5.0e-08
12	4096	4.1e-11	3.3e-06	2.8e-11	2.2e-06
14	16384	5.3e-09	1.9e-03	2.6e-09	9.4e-04
16	65536	1.0e-07	2.1e-01	3.9e-08	6.3e-02

the zeros of the polynomial $T_p(\lambda) - \sigma = 0$, where $p = n/2$ and $\sigma = \cos \alpha\pi$, denote

$$d_k = \begin{cases} b_1/2, & \text{if } k = 2p, \\ b_{2p-k+1}, & \text{if } p+1 \le k < p, \\ b_{p+1} - 2\sigma b_1, & \text{if } k = p, \\ b_{2p-k+1} - 4\sigma b_{p-k+1}, & \text{if } 1 \le k < p. \end{cases}$$

It follows from the discussion in section 4 that the polynomial $b(\lambda) = \sum_{k=1}^{2p} d_k U_{k-1}(\lambda)$. On the other hand, a simple calculation shows that

$$T_p'(t_j) = \frac{(-1)^{j-1} p \sin \alpha\pi}{\sin((j-\alpha_j)\pi/p)},$$

$$T_p''(t_j) = \frac{p((-1)^{j-1} \cot((j-\alpha_j)\pi/p) \sin \alpha\pi - p\sigma)}{\sin^2((j-\alpha_j)\pi/p)},$$

$$U_{k-1}(t_j) = \frac{\sin(k(j-\alpha_j)\pi/p)}{\sin((j-\alpha_j)\pi/p)},$$

$$U_{k-1}'(t_j) = \frac{(\cot((j-\alpha_j)\pi/p) \sin(k(j-\alpha_j)\pi/p) - k\cos(k(j-\alpha_j)\pi/p))}{\sin^2((j-\alpha_j)\pi/p)}.$$

It turns out that the solution (27), (28) becomes

$$x_{2j-1} = \frac{(-1)^{j-1}\sigma}{p\sin^3\alpha\pi}\sum_{k=1}^{2p} d_k \sin\frac{k(j-\alpha_j)}{p} - \frac{1}{p^2\sin^2\alpha\pi}\sum_{k=1}^{2p} kd_k\cos\frac{k(j-\alpha_j)}{p},$$
$$x_{2j} = \frac{\sin((j-\alpha_k)\pi/p)}{p^2\sin^2\alpha}\sum_{k=1}^{2p} d_k\sin\frac{k(j-\alpha_j)}{p}.$$

For a real x, define two p-vectors $\mathbf{u}(x) = (u_1,\ldots,u_p)^T$ and $\mathbf{v}(x) = (v_1,\ldots,v_p)^T$ by

$$u_{t+1} = \sum_{k=1}^{2p} d_k \sin\frac{k(2t+x)\pi}{p}, \qquad v_{t+1} = \sum_{k=1}^{2p} kd_k\cos\frac{k(2t+x)\pi}{p}.$$

Denote $\sigma(x) = \cos x\pi$. It follows from an elementary computation that

$$\begin{aligned}
u_{t+1} &= \sum_{k=1}^{2p} d_k\left(\sin\frac{2tk\pi}{p}\cos\frac{kx\pi}{p} + \cos\frac{2tk\pi}{p}\sin\frac{kx\pi}{p}\right)\\
&= \sum_{k=1}^{p}\left(d_k\cos\frac{kx\pi}{p} + d_{p+k}\cos\frac{(p+k)x\pi}{p}\right)\sin\frac{2tk\pi}{p}\\
&\quad + \left(d_k\sin\frac{kx\pi}{p} + d_{p+k}\sin\frac{(p+k)x\pi}{p}\right)\cos\frac{2tk\pi}{p}\\
&= \sum_{k=1}^{p}\left(d_k\cos\frac{kx\pi}{p} + \sigma(x)d_{p+k}\cos\frac{kx\pi}{p} - d_{p+k}\sin x\pi\sin\frac{kx\pi}{p}\right)\sin\frac{2kt\pi}{p}\\
&\quad + \sum_{k=1}^{p}\left(d_k\sin\frac{kx\pi}{p} + \sigma(x)d_{p+k}\sin\frac{kx\pi}{p} + d_{p+k}\sin x\pi\cos\frac{kx\pi}{p}\right)\cos\frac{2kt\pi}{p}\\
&= \sum_{k=1}^{p}(d_k + \sigma(x)d_{p+k})\left(\sin\frac{2kt\pi}{p}\cos\frac{kx\pi}{p} + \cos\frac{2kt\pi}{p}\sin\frac{kx\pi}{p}\right)\\
&\quad + \sin x\pi\sum_{k=1}^{p} d_{p+k}\left(-\sin\frac{2kt\pi}{p}\sin\frac{kx\pi}{p} + \cos\frac{2kt\pi}{p}\cos\frac{kx\pi}{p}\right).
\end{aligned}$$

To compute $\mathbf{u}(x)$ efficiently by FFT and make the statement clearly, let

$$\mathbf{c}_2 = (d_{2p}, d_{p+1},\ldots,d_{2p-1})^T, \qquad \mathbf{c}_1 = (d_p, d_1,\ldots,d_{p-1})^T + \sigma(x)\mathbf{c}_2.$$

Then the vector $\mathbf{u}(x)$ is easily represented by

$$\begin{aligned}
\mathbf{u}(x) &= -\sqrt{p}\,\mathrm{imag}\left(F_p\mathrm{diag}\left(\cos x\pi, \cos\frac{x\pi}{p},\ldots,\cos\frac{(p-1)x\pi}{p}\right)\mathbf{c}_1\right)\\
&\quad + \sqrt{p}\,\mathrm{real}\left(F_p\mathrm{diag}\left(\sin x\pi, \sin\frac{x\pi}{p},\ldots,\sin\frac{(p-1)x\pi}{p}\right)\mathbf{c}_1\right)\\
&\quad + \sqrt{p}\,\sin x\pi\,\mathrm{imag}\left(F_p\mathrm{diag}\left(\sin x\pi, \sin\frac{x\pi}{p},\ldots,\sin\frac{(p-1)x\pi}{p}\right)\mathbf{c}_2\right)\\
&\quad + \sqrt{p}\,\sin x\pi\,\mathrm{imag}\left(F_p\mathrm{diag}\left(\cos x\pi, \cos\frac{x\pi}{p},\ldots,\cos\frac{(p-1)x\pi}{p}\right)\mathbf{c}_2\right)
\end{aligned}$$

$$= -\sqrt{p}\,\mathrm{imag}\left(F_p\mathrm{diag}\left(\exp(-x\pi i), \exp\left(-\frac{x\pi i}{p}\right), \ldots, \exp\left(-\frac{(p-1)x\pi i}{p}\right)\right)\mathbf{c}_1\right)$$
$$+ \sqrt{p}\,\sin x\pi\,\mathrm{real}\left(F_p\left(\exp(x\pi i), \exp\left(\frac{x\pi i}{p}\right), \ldots, \exp\left(\frac{(p-1)x\pi i}{p}\right)\right)\mathbf{c}_2\right),$$

where $i = \sqrt{-1}$. Similarly, let $\mathbf{r}_2 = (2pd_{2p}, (p+1)d_{p+1}, \ldots, (2p-1)d_{2p-1})^T$ and $\mathbf{r}_1 = (pd_p, d_1, \ldots, (p-1)d_{p-1})^T + \sigma(x)\mathbf{r}_2$. We have

$$\mathbf{v}(x) = \sqrt{p}\,\mathrm{real}\left(F_p\mathrm{diag}\left(\exp(x\pi i), \exp\left(\frac{x\pi i}{p}\right), \ldots, \exp\left(\frac{(p-1)x\pi i}{p}\right)\right)\mathbf{r}_1\right)$$
$$+\sqrt{p}\,\sin x\pi\,\mathrm{imag}\left(F_p\left(\exp(-x\pi i), \exp\left(-\frac{x\pi i}{p}\right), \ldots, \exp\left(-\frac{(p-1)x\pi i}{p}\right)\right)\mathbf{r}_2\right).$$

Let $\mathbf{x}_{oo}$, $\mathbf{x}_{oe}$, $\mathbf{x}_{eo}$ and $\mathbf{x}_{ee}$ be subvectors of the solution $\mathbf{x}$ defined by

$$\mathbf{x}_{oo} = (x_1, x_5, \ldots, x_{4\lfloor\frac{p-1}{2}\rfloor+1})^T, \qquad \mathbf{x}_{oe} = (x_3, x_7, \ldots, x_{4\lfloor\frac{p}{2}\rfloor-1})^T,$$
$$\mathbf{x}_{eo} = (x_2, x_6, \ldots, x_{4\lfloor\frac{p-1}{2}\rfloor+2})^T, \qquad \mathbf{x}_{ee} = (x_1, x_5, \ldots, x_{4\lfloor\frac{p}{2}\rfloor})^T.$$

It is straightforward to show that

$$\mathbf{x}_{eo} = \frac{1}{p^2\sin^2\alpha\pi}\mathrm{diag}\left(\sin\frac{\alpha\pi}{p}, \ldots, \sin\frac{(2\lfloor\frac{p-1}{2}\rfloor+\alpha)\pi}{p}\right)(\mathbf{u}(\alpha))\left(1:\lceil\frac{p}{2}\rceil\right),$$
$$\mathbf{x}_{ee} = \frac{1}{p^2\sin^2\alpha\pi}\mathrm{diag}\left(\sin\frac{(2-\alpha)\pi}{p}, \ldots, \sin\frac{2\lfloor\frac{p}{2}\rfloor-\alpha)\pi}{p}\right)(\mathbf{u}(-\alpha))\left(2:\lceil\frac{p+1}{2}\rceil\right),$$
$$\mathbf{x}_{oo} = \frac{\sigma}{p\sin^3\alpha\pi}(\mathbf{u}(\alpha))\left(1:\lceil\frac{p}{2}\rceil\right) - \frac{1}{p^2\sin^2\alpha\pi}(\mathbf{v}(\alpha))\left(1:\lceil\frac{p}{2}\rceil\right),$$
$$\mathbf{x}_{oe} = -\frac{\sigma}{p\sin^3\alpha\pi}(\mathbf{u}(-\alpha))\left(2:\lceil\frac{p+1}{2}\rceil\right) - \frac{1}{p^2\sin^2\alpha\pi}(\mathbf{v}(-\alpha))\left(2:\lceil\frac{p+1}{2}\rceil\right).$$

Problem 2 of $V_{c,u}\mathbf{x} = \mathbf{b}$ and Problems 1 and 3 of $V_{c,t}\mathbf{x} = \mathbf{b}$ and $V_{c,u}\mathbf{x} = \mathbf{b}$ are implemented in similar ways. We omit further details.

Acknowledgments. The author is grateful to Gene Golub for helpful suggestions and to Nick Higham for valuable comments that helped to improve the presentation of the paper.

REFERENCES

[1] 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.

[2] C. Ballester and V. Pereyra, *On the construction of discrete approximations to linear differential expressions*, Math. Comp., 21 (1967), pp. 297–302.

[3] Å. Björck and T. Elfving, *Algorithms for confluent Vandermonde systems*, Numer. Math., 21 (1973), pp. 130–137.

[4] Å. Björck and V. Pereyra, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.

[5] P. J. Davis, *Circulant Matrices*, John Wiley, New York, 1979.

[6] N. Gastinel, *Inversion d'une matrice generalisant la matrice de hilbert*, Chiffres, 3 (1960), pp. 149–152.

[7] W. Gautschi, *The condition of Vandermonde-like matrices involving orthogonal polynomials*, Linear Algebra Appl., 52/53 (1983), pp. 293–300.
[8] A. Gerasoulis, *A fast algorithm for the multiplication of generalized Hilbert matrices with vectors*, Math. Comp., 50 (1988), pp. 179–188.
[9] A. Gerasoulis, M. D. Grigoriadis, and L. Sun, *A fast algorithm for Trummer's problem*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. s135–s138.
[10] I. Gohberg and V. Olshevsky, *Fast inversion of Chebyshev-Vandermonde matrices*, Numer. Math., 67 (1994), pp. 71–92.
[11] G. H. Golub, *Trummer's problem*, SIGACT News, ACM Special Interest Group on Automata and Computability Theory, 17 (1985), pp. 17.2–17.
[12] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulations*, J. Comput. Phys., 73 (1987), pp. 325–348.
[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] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, in Linear Algebra for Signal Processing, A. Bojanczyk and G. Cybenko, eds., IMA Volume in Mathematics and Its Application 69, Springer-Verlag, New York, 1994, pp. 63–81.
[15] N. J. Higham, *Error analysis of the Björck-Pereyra algorithm for solving Vandermonde systems*, Numer. Math., 50 (1987), pp. 613–632.
[16] N. J. Higham, *Fast solution of Vandermonde-like systems involving orthogonal polynomials*, IMA J. Numer. Anal., 8 (1988), pp. 473–486.
[17] N. J. Higham, *Stability analysis of algorithm for solving confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 23–41.
[18] N. J. Higham, *Iterative refinement enhances the stability of QR factorization methods for solving linear equations*, BIT, 31 (1991), pp. 447–468.
[19] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[20] J. Kautsky and S. Elhay, *Calculation of weights of interpolatory quadratures*, Numer. Math, 40 (1982), pp. 407–422.
[21] H. Lu, *Computational complexity of Vandermonde linear systems*, Chinese Sci. Bull., 9 (1990), pp. 654–656 (in Chinese).
[22] H. Lu, *On computational complexity of the multiplication of generalized Hilbert matrices with vectors and solution of certain generalized Hilbert linear systems*, Chinese Sci. Bull., 35 (1990), pp. 974–978.
[23] H. Lu, *Fast solution of confluent Vandermonde linear systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1277–1289.
[24] H. Lu, *Fast algorithms for confluent Vandermonde linear systems and generalized Trummer's problem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 655–674.
[25] H. Lu, *Solution of Vandermonde-like systems and confluent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl, (1996), pp. 127–138.
[26] 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.
[27] L. Reichel, *A matrix problem with application to rapid solution of integral equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 263–280.
[28] V. Rokhlin, *Rapid solution of integral equations of classical potential theory*, J. Comput. Phys., 60 (1985), pp. 187–207.
[29] J. F. Traub, *Associated polynomials and uniform methods for the solution of linear problems*, SIAM Rev., 8 (1966), pp. 277–301.
[30] M. Trummer, *An efficient implementation of a conformal mapping method using the Szegö kernel*, SIAM J. Numer. Anal., 23 (1986), pp. 853–872.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 1, pp. 277–278, January 1998

018

NOTE ON "FURTHER STUDY AND GENERALIZATION OF KAHAN'S MATRIX EXTENSION THEOREM"*

DAO-SHENG ZHENG†

Key words. Hermitian transformation, general solution of extension theorem, pseudoinverse form of solution

AMS subject classifications. 47A20, 15A09

PII. S0895479896314431

In the paper (see [5]) we quote the following theorem due to Kahan [6], [4, p. 232]: *Let* $R = (H^T, B^T)^T$, *where* H *is Hermitian. There exists a* W *such that the extended matrix* $A = \begin{pmatrix} H & B^* \\ B & W \end{pmatrix} = A^*$ *satisfies* $\| A \|_2 = \| R \|_2 = \rho$.

We showed in Theorem 4.1 of [5] the following general solution formula of W for Kahan's theorem:

$$B(\rho I + H)^+ B^* - \rho I \le W \le \rho I - B(\rho I - H)^+ B^*, \tag{1}$$

where the Moore–Penrose inverse of a matrix A is designated as A^+.

After publication of [5] it was pointed out that (1) is similar to a result of Krein in [2], [3] and that extensions of this result were also obtained in a later paper of Davis, Kahan, and Weinberger [1]. Theorem 1 of [2, p. 492] is as follows: *Let* A *be an Hermitian transformation with closed domain* $\mathcal{D}(A) \neq \mathcal{H}$ *and* $\| A \| \le 1$. *The set* $\mathcal{B}(A)$ *of all self-adjoint extensions* $\tilde{A}$ *of* A *with* $\| \tilde{A} \| \le 1$ *is nonvoid*; *moreover, it contains two transformations* A_μ *and* A_M ($A_\mu \le A_M$) *such that a bounded* $B \in \mathcal{B}(A)$ *if and only if* $A_\mu \le B \le A_M$. In the finite dimensional case, Theorem 4.1 of [5] is thus similar to Theorem 1 of [2] and Corollaries 1.3 and 1.4 of [1].

We were, of course, unaware of [1], [2], [3] at the time [5] was published. Also, the proofs are very different from each other, and the generalized inverse form (1) is new. Combining (1) with Lemma 2.7 of [5], one can easily identify the sensitivity of the data error of R to W.

From Theorem 4.1 of [5], the positive definite and semipositive definite extensions are discussed, respectively (see Theorems 5.1 and 5.2 of [5]). Also, based on Theorem 4.1 of [5], one can easily consider the problem of constructing a Hermitian extension W of a non-Hermitian matrix H and $R = (H^T, B^T)^T$.

REFERENCES

[1] C. DAVIS, W. M. KAHAN, AND H. F. WEINBERGER, *Norm-preserving dilations and their applications to optimal error bounds*, SIAM J. Numer. Anal., 19 (1982), pp. 445–469.

[2] M. G. KREIN, *The theory of self-adjoint extensions of semi-bounded Hermitian transformations and its applications*, Mat. Sb., 20 (1947), pp. 431–495; 21 (1947), pp. 365–404 (in Russian).

[3] M. G. KREIN, *On the trace formula in perturbation theory*, Mat. Sb., 33 (1953), pp. 597–626 (in Russian).

* Received by the editors December 20, 1996; accepted for publication (in revised form) by P. Van Dooren September 5, 1997.

http://www.siam.org/journals/simax/19-1/31443.html

† Department of Mathematics, East China Normal University, Shanghai 200062, People's Republic of China (cshen%fudan.edu.cn@ns.fudan.edu.cn).

[4] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980, p. 232.
[5] D. S. Zheng, *Further study and generalization of Kahan's matrix extension theorem*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 621–631.
[6] W. M. Kahan, *Inclusion Theorems for Clusters Eigenvalues of Hermitian Matrices*, Computer Science report, University of Toronto, Toronto, Canada, 1967.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 279–306, April 1998

001

STABLE AND EFFICIENT ALGORITHMS FOR STRUCTURED SYSTEMS OF LINEAR EQUATIONS*

MING GU†

Abstract. Recent research shows that structured matrices such as Toeplitz and Hankel matrices can be transformed into a different class of structured matrices called Cauchy-like matrices using the FFT or other trigonometric transforms. Gohberg, Kailath, and Olshevsky [*Math. Comp.*, 64 (1995), pp. 1557–1576] demonstrate numerically that their fast variation of the straightforward Gaussian elimination with partial pivoting (GEPP) procedure on Cauchy-like matrices is numerically stable. Sweet and Brent [*Adv. Signal Proc. Algorithms*, 2363 (1995), pp. 266–280] show that the error growth in this variation could be much larger than would be encountered with straightforward GEPP in certain cases. In this paper, we present a modified algorithm that avoids such extra error growth and can perform a fast variation of Gaussian elimination with complete pivoting (GECP). Our analysis shows that it is both efficient and numerically stable, provided that the element growth in the computed factorization is not large. We also present a more efficient variation of this algorithm and discuss implementation techniques that further reduce execution time. Our numerical experiments show that this variation is highly efficient and numerically stable.

Key words. displacement equation, error analysis, fast algorithm, Toeplitz matrix

AMS subject classifications. 15A06, 65F05, 65G05

PII. S0895479895291273

1. Introduction.

The Sylvester-type *displacement equation* for a matrix $M \in \mathbf{R}^{n\times n}$ is

$$\Omega \cdot M - M \cdot \Lambda = G, \tag{1.1}$$

where Ω and $\Lambda \in \mathbf{R}^{n\times n}$, and $G = A \cdot B$ with $A \in \mathbf{R}^{n\times\alpha}$ and $B \in \mathbf{R}^{\alpha\times n}$. The matrix pair (A, B) (or the matrix G) is the *generator* of M with respect to Ω and Λ, $\alpha \le n$ is the *displacement rank* with respect to Ω and Λ if $\operatorname{rank}(G) = \alpha$, and M is considered to possess a *displacement structure* with respect to Ω and Λ if $\alpha \ll n$. Such displacement equations first appeared in [19], and the concept of displacement structure was first introduced in [21]. The most general form of displacement structure, which includes (1.1) as a special case, was introduced in [22].

1.1. Fast algorithms for structured matrices.

The coefficient matrices in many linear systems of equations arising from signal processing, control theory, and interpolation applications often have such displacement structures. For example, the Cauchy-like matrix is a matrix of the following form (see [11, 17]):

$$C = \left(\frac{a_i^T \cdot b_j}{\omega_i - \lambda_j}\right)_{1\le i,j\le n} \qquad (a_i, b_j \in \mathbf{R}^\alpha),$$

*Received by the editors September 5, 1995; accepted for publication (in revised form) by Y. Genin December 16, 1996. This work was supported in part by the Applied Mathematical Sciences Subprogram of the Office of Energy Research, U.S. Department of Energy, under contract DE-AC03-76SF00098.

http://www.siam.org/journals/simax/19-2/29127.html

†Department of Mathematics, University of California, Los Angeles, CA 90095-1555 (mgu@math.ucla.edu).

where we assume that $\omega_i \neq \lambda_j$ for $1 \leq i, j \leq n$. Equivalently, we can define a Cauchy-like matrix to be the unique solution to the displacement equation

$$\Omega \cdot C - C \cdot \Lambda = A \cdot B$$

with

$$\Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n),\ \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n), \text{and}\ \ A = \begin{pmatrix} a_1^T \\ \vdots \\ a_n^T \end{pmatrix},\ B = (b_1, \ldots, b_n)\,.$$

In particular, C is a Cauchy matrix if $a_i^T \cdot b_j = 1$ for all i and j. We note that while the rank of C can be as large as n, its displacement rank is at most α.

Other classes of structured matrices include the Toeplitz matrices and the Hankel matrices. A Toeplitz matrix T is a matrix whose entries are constant along every diagonal ($T = (t_{i-j})_{1 \leq i,j \leq n}$), and a Hankel matrix H is a matrix whose entries are constant along every antidiagonal ($H = (h_{i+j-2})_{1 \leq i,j \leq n}$). These two classes of matrices are included in the larger class of Toeplitz-plus-Hankel matrices. A Toeplitz-plus-Hankel matrix is the sum of a Toeplitz and a Hankel matrix.

There are many fast algorithms that solve the Toeplitz (or Hankel, or Toeplitz-plus-Hankel) system of linear equations in $O(n^2)$ floating point operations, as opposed to $O(n^3)$ floating point operations normally required for a general dense matrix; there are also superfast algorithms that require only $O(n \log_2^2 n)$ floating point operations. However, all these fast and superfast algorithms are in general numerically unstable for indefinite systems. For discussions of some of these methods, see [3, 8, 25] and the references therein. Attempts to overcome this numerical instability using look-ahead techniques result in algorithms that could require (n^3) floating point operations in the worst case [4, 16].

Recently, Fiedler [10], Gohberg and Olshevsky [12], and Pan [24] show that Toeplitz and Hankel matrices can be tranformed into Cauchy-like matrices, and Gohberg and Olshevsky [14] present a fast variation (requiring only $O(n^2)$ floating point operations) of the straightforward GEPP procedure to solve a Cauchy-like linear system of equations. Heinig [17] is the first to solve the Toeplitz linear system of equations by transforming the Toeplitz matrix into a Cauchy-like matrix via fast Fourier or trigonometric transforms, and then solving the Cauchy-like linear system of equations via a fast variation of the straightforward GEPP. Among other results, Gohberg, Kailath, and Olshevsky [11] develop an improved version, Algorithm GKO, of Heinig's algorithm and demonstrate numerically that it is stable. They also transform the Hankel matrix and the Toeplitz-plus-Hankel matrix via fast trigonometric transforms into Cauchy-like matrices.

Sweet and Brent [26] have done an error analysis for the algorithms of [11]. They show that the error propagation of Algorithm GKO depends not only on the magnitudes of the triangular factors L and U in the LU factorization of the corresponding Cauchy-like matrix but also on the generator for this Cauchy-like matrix. They show that in some cases the generator can suffer large internal element growth and cause a corresponding growth in the backward and forward error; their results imply that Algorithm GKO is less numerically stable than the straightforward GEPP on the Cauchy-like matrix.

1.2. Main results. In this paper, we show how to avoid such internal element growth in the generator when factorizing the Cauchy-like matrix; we demonstrate

how to triangular factorize this Cauchy-like matrix using a variation of GECP in $O(n^2)$ floating point operations (see section 2). We compare a different choice of displacement equation for the Toeplitz and Toeplitz-plus-Hankel matrices with those in [11, 17] in terms of efficiency and numerical accuracy in factorizing the resulting Cauchy-like matrix, and based on our analysis and with this choice, we provide a new algorithm for factorizing a Toeplitz or Toeplitz-plus-Hankel matrix (see section 3) that performs about 50% less floating point operations than Algorithm GKO of [11]. We report interesting numerical experiments with this new algorithm (see section 4). And we perform an error analysis for fast Cauchy-like matrix factorization algorithms and show that this new algorithm is numerically stable, provided that the magnitude of the triangular factor U in the LU factorization is not large (see section 5).

We also discuss some implementation techniques that significantly reduce the amount of memory traffic during the execution of this new algorithm. Our numerical experiments indicate that they make the new algorithm up to a factor of 2 faster (see section 4).

1.3. Overview. In section 2 we review the fast algorithm of [11] for Cauchy-like matrices; we present a fast algorithm, Algorithm 2, that performs a variation of GECP on such matrices and avoids internal element growth in the generator; and we provide a variation of Algorithm 2 that is more efficient. In section 3 we compare different choices of displacement equation for the Toeplitz and Toeplitz-plus-Hankel matrices in terms of efficiency and numerical accuracy in factorizing the resulting Cauchy-like matrix; based on Algorithm 2 and a new choice of displacement equation, we provide a new algorithm, Algorithm 4, for solving the Toeplitz and Toeplitz-plus-Hankel system of linear equations. In section 4 we present numerical experiments with Algorithm 4 and compare this algorithm with some other available algorithms. In section 5 we perform an error analysis for Algorithms 2 and 4. And in section 6 we discuss some extensions, draw conclusions, and discuss some open problems.

1.4. Notation and conventions. For a matrix A, $|A|$ is the matrix of moduli of the $\{a_{i,j}\}$; $A_{p:q,s:k}$ is a submatrix of A that selects rows p to q of columns s to k; $A_{:,s:k}$ and $A_{s:k,:}$ select sth through kth rows and columns, respectively; and when $s = k$, we replace $s : k$ by s. Without loss of generality we assume A to be real unless it is specified to be complex. Our discussion for real matrices generally carries over to the complex case.

We will use the *max norm*, the ∞-*norm*, and the 2-*norm*

$$\|A\|_{\max} = \max_{i,j} |a_{i,j}|,\ \|A\|_\infty = \max_i \sum_j |a_{i,j}|,\ \|A\|_2 = \max_{\|u\|_2=1} \|A \cdot u\|_2,$$

as well as the *Frobenius norm* $\|A\|_F = \sqrt{\sum_{i,j} |a_{i,j}|^2}$. For a matrix $A \in \mathbf{R}^{n\times m}$, the following inequalities hold:

$$\frac{\|A\|_F}{\sqrt{n \cdot m}} \le \|A\|_{\max} \le \|A\|_2 \quad \text{and} \quad \frac{\|A\|_2}{\sqrt{n}} \le \|A\|_\infty \le \sqrt{m} \cdot \|A\|_2 \,. \tag{1.2}$$

P is a permutation matrix, and $P(j, k)$ denotes the permutation that interchanges the jth and kth rows of a matrix.

ϵ is the machine precision, and n is the order of the matrix to be factorized.

A *flop* is a floating point operation $x \circ y$, where x and y are floating point numbers and $\circ$ is one of $+$, $-$, $\times$, and $\div$. Taking the absolute value or comparing two floating point numbers is also counted as a flop.

In our error analysis, we take the usual model of arithmetic:[1]

$$\mathbf{fl}(x \circ y) = (x \circ y)(1 + \eta), \tag{1.3}$$

where $\mathbf{fl}(x \circ y)$ is the floating point result of the operation $\circ$ and $|\eta| \leq \epsilon$. For simplicity, we ignore the possibility of overflow and underflow.

Let $\bar{A} = A - a \cdot b^T$, where A is a matrix and a and b are vectors, $\mathbf{fl}(\bar{A})$ is the result of computing $\bar{A}$ in finite precision.

2. Gaussian elimination with pivoting for Cauchy-like matrices. Given a matrix $C \in \mathbf{R}^{n \times n}$, the first step of Gaussian elimination is to zero-out the first column of C below the diagonal entry:

$$C = \begin{pmatrix} \gamma_1 & u^T \\ r & \tilde{C}_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ l & I \end{pmatrix} \cdot \begin{pmatrix} \gamma_1 & u^T \\ 0 & C^{(2)} \end{pmatrix}, \tag{2.1}$$

where γ_1 is the pivot, $l = r/\gamma_1$, and $C^{(2)} = \tilde{C}_2 - l \cdot u^T$ is the Schur complement of γ_1. Gaussian elimination then recursively applies this step to $C^{(2)}$. At the end of this procedure, C is factored into $C = L \cdot U$, where L is a lower triangular matrix and U is an upper triangular matrix.

The following theorem shows that if C is a Cauchy-like matrix with displacement rank α, so is $C^{(2)}$. The algorithms of Gohberg, Kailath, and Olshevsky [11] are based on this theorem. More general forms of it appear in [11, 13, 14, 22], and a variation of it appears in [17].

THEOREM 2.1. *Let matrix C in* (2.1) *satisfy the displacement equation*

$$\Omega \cdot C - C \cdot \Lambda = A \cdot B \tag{2.2}$$

with $\Omega = \operatorname{diag}(\omega_1, \Omega_2)$ and $\Lambda = \operatorname{diag}(\lambda_1, \Lambda_2) \in \mathbf{R}^{n \times n}$ diagonal, $A = \binom{a_1^T}{\tilde{A}_2} \in \mathbf{R}^{n \times \alpha}$, and $B = (b_1 \ \tilde{B}_2) \in \mathbf{R}^{\alpha \times n}$. Assume that $\gamma_1 \neq 0$. Then $C^{(2)}$ satisfies the displacement equation

$$\Omega_2 \cdot C^{(2)} - C^{(2)} \cdot \Lambda_2 = A^{(2)} \cdot B^{(2)} \tag{2.3}$$

with $A^{(2)} = \tilde{A}_2 - l \cdot a_1^T \in \mathbf{R}^{(n-1) \times \alpha}$ and $B^{(2)} = \tilde{B}_2 - b_1 \cdot u^T/\gamma_1 \in \mathbf{R}^{\alpha \times (n-1)}$.

Hence one step of Gaussian elimination on C involves computing the first row and column γ_1, r, and u of C from (2.2) and computing the vector l. To recursively apply this procedure to $C^{(2)}$, its generator $(A^{(2)}, B^{(2)})$ is then computed from (2.3).

2.1. Partial pivoting. Partial pivoting is a strategy to reduce the element growth in the LU factorization. To perform partial pivoting on the first column of C, one finds its largest magnitude entry $(k_{\max}, 1)$, permute it to the $(1, 1)$ entry to get $P(1, k_{\max}) \cdot C$, and then applies one elimination step to $P(1, k_{\max}) \cdot C$. Let C be a Cauchy-like matrix satisfying (2.2). Then for every k

$$(P(1,k) \cdot \Omega \cdot P(1,k)) \cdot (P(1,k) \cdot C) - (P(1,k) \cdot C) \cdot \Lambda = (P(1,k) \cdot A) \cdot B,$$

where $(P(1,k) \cdot \Omega \cdot P(1,k))$ is again a diagonal matrix. In particular, this implies that $P(1, k_{\max}) \cdot C$ is a Cauchy-like matrix. Algorithm 1 below performs fast GEPP for a

[1] This model excludes some CRAY machines that do not have a guard digit. Our error analysis still holds for such machines with a few easy modifications.

Cauchy-like matrix C. It is suggested by Gohberg, Kailath, and Olshevsky [11]. The recursions for computing A and B (without explicitly computing L and U) and the partial pivoting idea are from [17].

ALGORITHM 1. Fast GEPP for a Cauchy-like matrix.

$L := 0$; $U := 0$; $P := I$;
for $k = 1$ to n **do**
  $L_{k:n,k} := (\Omega_{k:n,k:n} - \lambda_k I)^{-1} \cdot A_{k:n,:} \cdot B_{:,k}$;
  $k_{\max} := \operatorname{argmax}_{k \le j \le n} |L_{j,k}|$;
  if $k_{\max} > k$ **then**
    $P := P \cdot P(k, k_{\max})$; $\Omega := P(k, k_{\max}) \cdot \Omega \cdot P(k, k_{\max})$;
    $A = P(k, k_{\max}) \cdot A$; $L_{:,1:k} = P(k, k_{\max}) \cdot L_{:,1:k}$;
  endif
  $U_{k,k} = L_{k,k}$; $U_{k,k+1:n} := A_{k,:} \cdot B_{:,k+1:n} \cdot (\omega_k I - \Lambda_{k+1:n,k+1:n})^{-1}$;
  $L_{k,k} = 1$; $L_{k+1:n,k} := L_{k+1:n,k}/U_{k,k}$;
  $A_{k+1:n,:} = A_{k+1:n,:} - L_{k+1:n,k} \cdot A_{k,:}$; $B_{:,k+1:n} = B_{:,k+1:n} - B_{:,k+1} \cdot U_{k,k+1:n}/U_{k,k}$.
endfor

Remark 1. If the input data A, B, Ω, and Λ are real, Algorithm 1 costs about $(4\alpha+2.5)n^2$ flops; there is also potentially about $n^2/2$ swaps of memory locations. For a matrix transformed into a Cauchy-like matrix from a Toeplitz-plus-Hankel matrix (see section 1 and section 3), the displacement rank α is at most 4. In this case, Algorithm 1 costs about $18.5n^2$ flops.

Remark 2. If the input data are complex, Algorithm 1 costs about $(16\alpha + 12)n^2$ flops[2]; there is also potentially about n^2 swaps of memory locations. For a matrix transformed into a Cauchy-like matrix from a Toeplitz matrix (see section 3), the displacement rank α is at most 2. In this case, Algorithm 1 costs about $44n^2$ flops.

We observe that Algorithm 1 produces the same LU factorization as that of straightforward GEPP on C. Hence, one potential problem with Algorithm 1 is the *element growth* in the LU factorization. Let U be the upper triangular matrix in the LU factorization, and let $g_{PP} \equiv \|U\|_{\max}/\|C\|_{\max}$ be the *element growth factor*. It is well known that $g_{PP} \le 2^{n-1}$ for GEPP, and although very rare, this bound is attainable for certain dense matrices [15, pp. 115–116]. It is not clear whether this bound is attainable for Cauchy-like matrices with low displacement rank. When large element growth does occur, the computed LU factorizations can be very inaccurate.

2.2. Complete pivoting. Complete pivoting may in general further reduce element growth in the LU factorization. To perform complete pivoting on C, one finds its largest magnitude entry $(k_{\max}, j_{\max})$ in the entire matrix, permute it to the $(1,1)$ entry to get $P(1, k_{\max}) \cdot C \cdot P(1, j_{\max})$, and then applies the elimination step to this permuted matrix. Let C be a Cauchy-like matrix satisfying equation (2.2). Then for every $1 \le k, j \le n$

$$\begin{aligned}&(P(1,k) \cdot \Omega \cdot P(1,k)) \cdot (P(1,k) \cdot C \cdot P(1,j)) \\ -\,&(P(1,k) \cdot C \cdot P(1,j)) \cdot (P(1,j) \cdot \Lambda \cdot P(1,j)) = (P(1,k) \cdot A) \cdot (B \cdot P(1,j))\,.\end{aligned}$$

In particular, this equation implies that $P_{1,k_{\max}} \cdot C \cdot P_{1,j_{\max}}$ is still a Cauchy-like matrix.

[2]We count a complex addition or subtraction as 2 flops, a complex multiplication as 6 flops, a complex division as 10 flops, and the total cost of taking absolute value and performing comparison as 4 flops.

However, finding the largest magnitude entry $(k_{\max}, j_{\max})$ of C costs $O(n^2)$ flops in general. If this is done on every step of Gaussian elimination, then the total cost will be $O(n^3)$, which is too expensive.

On the other hand, it is not absolutely necessary to use *the* largest magnitude entry as pivot in order to reduce element growth. Any entry sufficiently large in magnitude should do.

Define

$$\xi_{\max} = \max_{1 \le i,j \le n} |\omega_i - \lambda_j| \ , \ \xi_{\min} = \min_{1 \le i,j \le n} |\omega_i - \lambda_j| \ , \text{ and } \rho = \frac{\xi_{\max}}{\xi_{\min}}. \tag{2.4}$$

The following lemma tells us where to look for such an entry.

LEMMA 2.2. *Let C be a Cauchy-like matrix satisfying* (2.2)*, and let the* $j\text{th}_{\max}$ *column be the largest* 2*-norm column of $A \cdot B$. Then*

$$\|C\|_{\max} \le \sqrt{n} \cdot \rho \cdot \|C_{:,j_{\max}}\|_\infty \quad \textit{and} \quad \|C\|_F \le n \cdot \rho \cdot \|C_{:,j_{\max}}\|_\infty.$$

Proof. Let $|G_{i_G,j_{\max}}| = \|G_{:,j_{\max}}\|_\infty$, where $G = A \cdot B$. Then for any $1 \le s, j \le n$

$$\begin{aligned} |C_{s,j}| &= \frac{|G_{s,j}|}{|\omega_s - \lambda_j|} \le \frac{\|G_{:,j}\|_2}{|\omega_s - \lambda_j|} \le \frac{\|G_{:,j_{\max}}\|_2}{|\omega_s - \lambda_j|} \\ &\le \frac{\sqrt{n} \cdot |G_{i_G,j_{\max}}|}{|\omega_s - \lambda_j|} = \frac{\sqrt{n} \cdot |\omega_{i_G} - \lambda_{j_{\max}}|}{|\omega_s - \lambda_j|} \cdot |C_{i_G,j_{\max}}| \\ &\le \frac{\sqrt{n} \cdot \xi_{\max}}{\xi_{\min}} \cdot \|C_{:,j_{\max}}\|_\infty. \end{aligned}$$

Hence, the first assertion of the lemma follows immediately.

For the second assertion we have

$$\begin{aligned} \|C\|_F^2 &= \sum_{s,j} |C_{s,j}|^2 = \sum_{s,j} \frac{|G_{s,j}|^2}{|\omega_s - \lambda_j|^2} \le \frac{\sum_{s,j} |G_{s,j}|^2}{\xi_{\min}^2} \\ &\le \frac{n \cdot \sum_s |G_{s,j_{\max}}|^2}{\xi_{\min}^2} \le \frac{n \cdot \sum_s |C_{s,j_{\max}}|^2 \cdot \xi_{\max}^2}{\xi_{\min}^2} \\ &\le \frac{n^2 \cdot \|C_{:,j_{\max}}\|_\infty^2 \cdot \xi_{\max}^2}{\xi_{\min}^2}. \end{aligned}$$

To finish the proof, we take square roots on both sides. ∎

To find the column $j_{\max}$ in Lemma 2.2, we QR factorize A to get $A = \mathcal{A} \cdot R$, where $\mathcal{A} \in \mathbf{R}^{n \times \alpha}$ is column orthogonal and R is upper triangular. We then compute $\mathcal{B} = R \cdot B$. It follows that

$$A \cdot B = \mathcal{A} \cdot \mathcal{B}. \tag{2.5}$$

Since $\mathcal{A}$ is column orthogonal, the jth columns of $A \cdot B$ and $\mathcal{B}$ have the same 2-norm for $1 \le j \le n$. Algorithm 2 below differs from Algorithm 1 in that we compute $j_{\max}$ by looking for the largest 2-norm column of $\mathcal{B}$ and we perform generator re-decomposition (2.5). Algorithm 2 assumes that the matrix A is column orthogonal on input.

ALGORITHM 2. Fast GECP for a Cauchy-like matrix.
$L := 0$; $U := 0$; $P := I$; $Q := I$;
for $k = 1$ to n **do**
  $j_{\max} := \operatorname{argmax}_{k \le j \le n} \|B_{:,j}\|_2$;
  if $j_{\max} > k$ **then**
    $Q := P(k, j_{\max}) \cdot Q$; $\Lambda := P(k, j_{\max}) \cdot \Lambda \cdot P(k, j_{\max})$;
    $B := B \cdot P(k, j_{\max})$; $U_{1:k,:} = U_{1:k,:} \cdot P(k, j_{\max})$;
  endif
  $L_{k:n,k} := (\Omega_{k:n,k:n} - \lambda_k I)^{-1} \cdot A_{k:n,:} \cdot B_{:,k}$;
  $k_{\max} := \operatorname{argmax}_{k \le j \le n} |L_{j,k}|$;
  if $k_{\max} > k$ **then**
    $P := P \cdot P(k, k_{\max})$; $\Omega := P(k, k_{\max}) \cdot \Omega \cdot P(k, k_{\max})$;
    $A := P(k, k_{\max}) \cdot A$; $L_{:,1:k} := P_{k,k_{\max}} \cdot L_{:,1:k}$;
  endif
  $U_{k,k} := L_{k,k}$; $U_{k,k+1:n} := A_{k,:} \cdot B_{:,k+1:n} \cdot (\omega_k I - \Lambda_{k+1:n,k+1:n})^{-1}$;
  $L_{k,k} := 1$; $L_{k+1:n,k} := L_{k+1:n,k}/U_{k,k}$;
  $A_{k+1:n,:} := A_{k+1:n,:} - L_{k+1:n,k} \cdot A_{k,:}$; $B_{:,k+1:n} := B_{:,k+1:n} - B_{:,k+1} \cdot U_{k,k+1:n}/U_{k,k}$;
  $A_{k+1:n,:} := \mathcal{A} \cdot R$ (QR factorization of $A_{k+1:n,:}$); $\mathcal{B} := R \cdot B_{:,k+1:n}$;
  $A_{k+1:n,:} := \mathcal{A}$, $B_{:,k+1:n} := \mathcal{B}$.
endfor

For the rest of section 2.2, we derive an upper bound on the element growth factor for Algorithm 2, using techniques similar to those used by Wilkinson [28] to bound the growth factor for the straightforward GECP. In section 2.3 we will discuss Algorithm 2 in more detail, and in section 5.5 we will show that Algorithm 2 is numerically stable provided that the U matrix is not large in norm.

Let

$$\mathcal{W}(k) = \left(k \prod_{s=2}^{k} s^{1/(s-1)} \right)^{1/2} = O\left(k^{\frac{1}{2} + \frac{1}{4} \ln k} \right),$$

which is Wilkinson's upper bound on the growth factor for GECP on a $k \times k$ matrix. Although $\mathcal{W}(k)$ is not a polynomial in k, it does not grow very fast either [28].

We will need the following well-known result.

LEMMA 2.3 (see Householder [20, p. 15]). *For any $C \in \mathbf{R}^{n \times n}$, we have*

$$|\det C| \le \left(\frac{\|C\|_F}{\sqrt{n}} \right)^n.$$

THEOREM 2.4. *Let C be a Cauchy-like matrix satisfying* (2.2)*, and let $C = P \cdot L \cdot U \cdot Q$ be the LU factorization generated by Algorithm* 2 *in exact arithmetic. Then the element growth factor $g_{CP} \equiv \|U\|_{\max} / \|C\|_{\max}$ satisfies*

$$g_{CP} \le \sqrt{n} \cdot \rho^{2 + \sum_{k=1}^{n-1} 1/k} \cdot \mathcal{W}(n). \tag{2.6}$$

Proof. Without loss of generality we assume that pivoting has been done before hand, so that Algorithm 2 does not perform any pivoting.

For $1 \le k \le n$, let $C^{(k)} \in \mathbf{R}^{(n-k+1)\times(n-k+1)}$ be the Cauchy-like matrix to be factored at the kth step in Algorithm 2, with γ_k being the pivot (the $(1,1)$ entry of $C^{(k)}$). We note that $C^{(1)} = C$ in this notation.

Since Algorithm 2 performs partial pivoting, we have $|\gamma_k| = \|C^{(k)}_{:,1}\|_\infty$, and since the first column of the generator for $C^{(k)}$ has the largest column 2-norm, we have $\|C^{(k)}\|_F \le (n-k+1)\cdot\rho\cdot|\gamma_k|$ according to Lemma 2.2. It follows from Lemma 2.3 that

$$\begin{aligned}\left|\det\left(C^{(k)}\right)\right| &\le \left(\frac{\|C^{(k)}\|_F}{\sqrt{n-k+1}}\right)^{n-k+1} \le \left(\frac{(n-k+1)\cdot\rho\cdot|\gamma_k|}{\sqrt{n-k+1}}\right)^{n-k+1} \\ &= \left(\sqrt{n-k+1}\cdot\rho\cdot|\gamma_k|\right)^{n-k+1}.\end{aligned}$$

On the other hand,

$$\left|\det\left(C^{(k)}\right)\right| = |\gamma_k|\cdots|\gamma_n|.$$

Comparing these two relations we have

$$|\gamma_k|\cdots|\gamma_n| \le \left(\rho\cdot\sqrt{n-k+1}\cdot|\gamma_k|\right)^{n-k+1}, \quad 1 \le k \le n. \tag{2.7}$$

Since

$$\sum_{k=1}^{s}\frac{1}{(n-k)(n-k+1)} + \frac{1}{n} = \frac{1}{n-s},$$

taking the product of the $(n-k)(n-k+1)$st root of (2.7) with $k = 1, 2, \ldots, n-1$ and the nth root of (2.7) with $k = 1$, we have

$$\begin{aligned}\prod_{s=1}^{n-1}|\gamma_s|^{1/(n-s)}\cdot\gamma_n &\le \left(\prod_{k=1}^{n-1}\left(\rho\sqrt{n-k+1}\cdot\gamma_k\right)^{1/(n-k)}\right)\cdot\left(\rho\sqrt{n}\cdot\gamma_1\right) \\ &= \rho^{1+\sum_{k=1}^{n-1}1/(n-k)}\cdot\left(n\cdot\prod_{k=1}^{n-1}(n-k+1)^{1/(n-k)}\right)^{1/2}\cdot\left(\prod_{k=1}^{n-1}|\gamma_k|^{1/(n-k)}\right)\cdot|\gamma_1|,\end{aligned}$$

which simplifies to

$$|\gamma_n| \le |\gamma_1|\cdot\rho^{1+\sum_{k=1}^{n-1}1/k}\cdot\left(n\cdot\prod_{k=2}^{n}k^{1/(k-1)}\right)^{1/2} = |\gamma_1|\cdot\rho^{1+\sum_{k=1}^{n-1}1/k}\cdot\mathcal{W}(n).$$

Repeating the same argument allows us to conclude that

$$|\gamma_s| \le |\gamma_1|\cdot\rho^{1+\sum_{k=1}^{n-1}1/k}\cdot\mathcal{W}(n), \quad 1 \le s \le n.$$

It now follows from Lemma 2.2 that

$$\|C^{(s)}\|_{\max} \le \sqrt{n}\cdot\rho\cdot|\gamma_s| \le |\gamma_1|\cdot\sqrt{n}\cdot\rho^{2+\sum_{k=1}^{n-1}1/k}\cdot\mathcal{W}(n).$$

To complete the proof, we observe that the sth row of the upper triangular matrix U is the first row of $C^{(s)}$. Hence,

$$\|U\|_{\max} \le |\gamma_1| \cdot \sqrt{n} \cdot \rho^{2+\sum_{k=1}^{n-1} 1/k} \cdot \mathcal{W}(n).$$

The assertion of the theorem follows immediately from $|\gamma_1| \le \|C\|_{\max}$. □

Remark 3. The determinant argument in the proof of Theorem 2.4 ignores the fact that C is a Cauchy-like matrix; hence, the upper bound provided by (2.6) could be much larger than necessary, especially for Cauchy-like matrices with low displacement rank.

Remark 4. Since $\sum_{k=1}^{n-1} 1/k = \ln n + O(1)$, the bound (2.6) simplifies to

$$g_{CP} \le n^{\ln \rho + \frac{1}{4} \ln n + O(1)}.$$

Assume that $\rho = O(n^\beta)$ for a constant β, then

$$g_{CP} \le n^{\frac{4\beta+1}{4} \ln n + O(1)}.$$

If C is transformed into a Cauchy-like matrix from a Toeplitz matrix (or a Toeplitz-plus-Hankel matrix) via any of the transforms discussed in section 3, then $1 \le \beta \le 3$. Although this upper bound is much larger than $\mathcal{W}(n)$, it is still much smaller than 2^{n-1}.

2.3. Further considerations. In addition to the potential element growth in the LU factorization, Sweet and Brent [26] show that the generator (A, B) updated as in Algorithm 1 could also grow so that

$$\||A_{k:n,:}| \cdot |B_{:,k:n}|\|_2 \gg \|A_{k:n,:} \cdot B_{:,k:n}\|_2$$

for some k. And if this happens, the backward and forward error could become large.

However, such element growth in the generator can easily be avoided. Since $A_{k:n,:}$ is kept column orthogonal for all k in Algorithm 2, it follows that

$$\begin{aligned}\||A_{k:n,:}| \cdot |B_{:,k:n}|\|_2 &\le \||A_{k:n,:}|\|_2 \cdot \||B_{:,k:n}|\|_2 \le \|A_{k:n,:}\|_F \cdot \|B_{:,k:n}\|_F \\ &\le \sqrt{\alpha} \cdot \|B_{:,k:n}\|_F \le \alpha \cdot \|B_{:,k:n}\|_2 = \alpha \cdot \|A_{k:n,:} \cdot B_{:,k:n}\|_2.\end{aligned}$$

Hence, keeping $A_{k:n,:}$ column orthogonal for all k also has the additional advantage of avoiding potential element growth within the generator (A, B). In fact, such growth can be avoided as long as $A_{k:n,:}$ is well conditioned.

From a practical point of view, it does not seem necessary to column orthogonalize $A_{k:n,:}$ at every step just to keep it well conditioned, nor does it seem necessary to perform pivoting on the columns at every step to reduce element growth. As a practical modification to Algorithm 2, the following algorithm performs these operations only once in every K steps, where K is a user-provided positive integer. It assumes that the matrix A is initially column orthogonal.

ALGORITHM 3. Practical modification to Algorithm 2.

$L := 0$; $U := 0$; $P := I$; $Q := I$;
for $k = 1$ to n **do**
  if (**mod**$(k, K) = 1$) **then**
    $j_{\max} := \text{argmax}_{k \le j \le n} \|B_{:,j}\|_2$;

if $j_{\max} > k$ **then**
$Q := P(k, j_{\max}) \cdot Q$; $\Lambda := P(k, j_{\max}) \cdot \Lambda \cdot P(k, j_{\max})$;
$B := B \cdot P(k, j_{\max})$; $U_{1:k,:} = U_{1:k,:} \cdot P(k, j_{\max})$;
endif
endif
$L_{k:n,k} := (\Omega_{k:n,k:n} - \lambda_k I)^{-1} \cdot A_{k:n,:} \cdot B_{:,k}$;
$k_{\max} := \text{argmax}_{k \le j \le n} |L_{j,k}|$;
if $k_{\max} > k$ **then**
$P := P \cdot P(k, k_{\max})$; $\Omega := P(k, k_{\max}) \cdot \Omega \cdot P(k, k_{\max})$;
$A := P(k, k_{\max}) \cdot A$; $L_{:,1:k} := P_{k,k_{\max}} \cdot L_{:,1:k}$;
endif
$U_{k,k} := L_{k,k}$; $U_{k,k+1:n} := A_{k,:} \cdot B_{:,k+1:n} \cdot (\omega_k I - \Lambda_{k+1:n,k+1:n})^{-1}$;
$L_{k,k} := 1$; $L_{k+1:n,k} := L_{k+1:n,k}/U_{k,k}$;
$A_{k+1:n,:} := A_{k+1:n,:} - L_{k+1:n,k} \cdot A_{k,:}$; $B_{:,k+1:n} := B_{:,k+1:n} - B_{:,k+1} \cdot U_{k,k+1:n}/U_{k,k}$;
if $(\mathbf{mod}(k, K) = 0)$ **then**
$A_{k+1:n,:} := \mathcal{A} \cdot R$ (QR factorization of $A_{k+1:n,:}$); $\mathcal{B} := R \cdot B_{:,k+1:n}$;
$A_{k+1:n,:} := \mathcal{A}$, $B_{:,k+1:n} := \mathcal{B}$.
endif
endfor

Remark 5. The cost for recomputing $A_{k+1:n,:}$ and $B_{:,k+1:n}$ through QR factorizations is about $5/2\alpha^2 n^2$ flops in real arithmetic and $10\alpha^2 n^2$ flops in complex arithmetic. However, if α is large and if QR factorization is performed at every step, these costs can be brought down to $O(\alpha n^2)$ by using QR updating techniques (see [15, section 12]). Our main interest in this paper is to use Algorithm 3 to factorize the Cauchy-like matrix that is transformed from a Toeplitz-plus-Hankel matrix (cf. section 1 and section 3). For such matrices α is at most 4. In our implementation, we recompute the QR factorization every $K = 10$ steps.

3. Factorizing Toeplitz-plus-Hankel-like matrices.

3.1. Factorizing Toeplitz-plus-Hankel-like matrices.

Define

$$Y_{\delta_1,\delta_2} = \begin{pmatrix} \delta_1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & \vdots \\ 0 & 1 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 & 1 \\ 0 & \cdots & 0 & 1 & \delta_2 \end{pmatrix}, \tag{3.1}$$

and $\Omega = Y_{1,1}$ and $\Lambda = Y_{1,-1}$. It is easy to verify that every Toeplitz-plus-Hankel matrix satisfies the displacement equation (1.1) with G having nonzero entries only in its first and last rows and columns, thus a matrix of rank at most 4. Hence, the displacement rank of a Toeplitz-plus-Hankel matrix is at most 4 (cf. [11, 18]). In particular, these results are true for every Toeplitz or Hankel matrix.

LEMMA 3.1. *Let $M \in \mathbf{R}^{n \times n}$ be a matrix satisfying the displacement equation*

$$Y_{1,1} \cdot M - M \cdot Y_{1,-1} = A \cdot B, \tag{3.2}$$

where $A \in \mathbf{R}^{n\times\alpha}$ *and* $B \in \mathbf{R}^{\alpha\times n}$. *Then* $\mathcal{Q}_{1,1}^T \cdot M \cdot \mathcal{Q}_{1,-1}$ *is a Cauchy-like matrix:*

$$\mathcal{D}_{1,1} \cdot \left(\mathcal{Q}_{1,1}^T \cdot M \cdot \mathcal{Q}_{1,-1}\right) - \left(\mathcal{Q}_{1,1}^T \cdot M \cdot \mathcal{Q}_{1,-1}\right) \cdot \mathcal{D}_{1,1} = \left(\mathcal{Q}_{1,1}^T \cdot A\right) \cdot \left(B \cdot \mathcal{Q}_{1,-1}\right), \tag{3.3}$$

where

$$\mathcal{Q}_{1,1} = \sqrt{\frac{2}{n}} \cdot \left(q_j \cos\frac{(2k-1)(j-1)\pi}{2n}\right)_{1\le k,j\le n}$$

and

$$\mathcal{Q}_{1,-1} = \left(\sqrt{\frac{2}{n}} \cdot \left(\cos\frac{(2k-1)(2j-1)\pi}{4n}\right)\right)_{1\le k,j\le n}$$

are orthogonal matrices with $q_1 = \frac{1}{2}$ *and* $q_j = 1$ *for* $2 \le j \le n$; *and*

$$\mathcal{D}_{1,1} = 2 \cdot \operatorname{diag}\left(1, \cos\frac{\pi}{n}, \ldots, \cos\frac{(n-1)\pi}{n}\right),$$
$$\mathcal{D}_{1,-1} = 2 \cdot \operatorname{diag}\left(\cos\frac{\pi}{2n}, \cos\frac{3\pi}{2n}, \ldots, \cos\frac{(2n-1)\pi}{2n}\right).$$

Proof. It can be checked that

$$Y_{1,1} = \mathcal{Q}_{1,1} \cdot \mathcal{D}_{1,1} \cdot \mathcal{Q}_{1,1}^T, \qquad Y_{1,-1} = \mathcal{Q}_{1,-1} \cdot \mathcal{D}_{1,-1} \cdot \mathcal{Q}_{1,-1}^T.$$

The lemma follows immediately by substituting the above equation into (3.2) and multiplying by $\mathcal{Q}_{1,1}^T$ from the left and by $\mathcal{Q}_{1,-1}$ from the right. ∎

We call a matrix M *Toeplitz-plus-Hankel-like* if it satisfies the displacement equation (3.2) with $\alpha \ll n$ (cf. [11]). To solve a Toeplitz-plus-Hankel-like linear system of equations

$$M \cdot x = z,$$

one can transform M into a Cauchy-like matrix using Lemma 3.1, factorize this matrix by using any of the methods discussed in section 2 to obtain a factorization of the form

$$M = \mathcal{Q}_{1,1}^T \cdot P \cdot L \cdot U \cdot Q \cdot \mathcal{Q}_{1,-1}^T, \tag{3.4}$$

and compute the solution to the linear system using this factorization. The idea of transforming a Toeplitz matrix into a Cauchy-like matrix was first proposed by Heinig [17], and the idea of transforming a Toeplitz-plus-Hankel matrix into a Cauchy-like matrix was first proposed by Gohberg, Kailath, and Olshevsky [11].

We summarize the above in Algorithm 4, assuming that M satisfies (3.2) with A column orthogonal.

ALGORITHM 4. Solving $M \cdot x = z$.

1. Set $\Omega := \mathcal{D}_{1,1}$; $\Lambda := \mathcal{D}_{1,-1}$; and compute $A := \mathcal{Q}_{1,1}^T \cdot A$; $B := B \cdot \mathcal{Q}_{1,-1}$.
2. Compute the factorization (3.4) by applying one of Algorithms 1, 2, and 3 with Ω, Λ, A, and B.
3. Compute $x = \mathcal{Q}_{1,-1} \cdot Q^T \cdot U^{-1} \cdot L^{-1} \cdot P^T \cdot \mathcal{Q}_{1,1} \cdot z$.

Both $\mathcal{Q}_{1,1}$ and $\mathcal{Q}_{1,-1}$ are fast trigonometric transform matrices; hence, the cost of step 1 is about $O(\alpha\, n \log_2 n)$ flops via 2α such transforms; similarly the cost of step 3 is about $2n^2$ flops via two fast trigonometric transforms, two permutations, and forward and backward substitution. The bulk of the cost is in step 2, factorizing the Cauchy-like matrix in (3.3).

For a real Toeplitz-plus-Hankel matrix, the displacement rank α is at most 4 in (3.3). When Algorithm 3 is used in step 2, the cost for step 2 is about $18.5n^2 + O(n^2/K)$ flops for a user-specified integer K (see Remark 5). Hence, Algorithm 4 takes about $20.5n^2 + O(n^2/K)$ flops to solve a Toeplitz-plus-Hankel system of equations. This is also true for a Toeplitz or a Hankel system of equations.

3.2. Comparison with previous methods. The Toeplitz-plus-Hankel matrix satisfies other displacement equations, too. It is known that [11, 18] for $\Omega = Y_{0,0}$ and $\Lambda = Y_{1,1}$ every Toeplitz-plus-Hankel matrix satisfies the displacement equation (1.1) with G having nonzero entries only in its first and last rows and columns. It is known that matrix $Y_{0,0}$ can be diagonalized using fast trigonometric transform matrices (see, for example, [2]):

$$Y_{0,0} = \mathcal{Q}_{0,0} \cdot \mathcal{D}_{0,0} \cdot \mathcal{Q}_{0,0}^T,$$

where $\mathcal{Q}_{0,0} = \frac{2}{n+1} \cdot (\sin \frac{kj\pi}{n+1})_{1 \le k,j \le n}$, and $\mathcal{D}_{0,0} = 2 \cdot \mathrm{diag}(\cos \frac{\pi}{n+1}, \ldots, \cos \frac{n\pi}{n+1})$. Gohberg, Kailath, and Olshevsky [11] suggest that to solve a Toeplitz-plus-Hankel system of linear equations, one transforms the coefficient matrix to a Cauchy-like matrix C that satisfies

$$\mathcal{D}_{0,0} \cdot C - C \cdot \mathcal{D}_{1,-1} = A \cdot B, \tag{3.5}$$

with $\mathrm{rank}(A), \mathrm{rank}(B) \le 4$, and one then applies Algorithm 1 to C. The resulting algorithm is named Algorithm TpH.

However, Algorithm TpH has some disadvantages over Algorithm 4. It can be shown that the parameter ρ defined in (2.4) is $O(n^3)$ for (3.5) and $O(n^2)$ for (3.3). Our upper bound on g_{CP} in section 2.2 and error analysis in section 5 suggest that the smaller ρ is, the smaller the potential element growth and backward error. Hence, Algorithm TpH could be potentially less accurate than Algorithm 4. Another disadvantage for Algorithm TpH is that in order for the fast trigonometric transforms with $\mathcal{Q}_{0,0}$ and $\mathcal{Q}_{1,1}$ to be very efficient, both n and $n+1$ must be products of small prime numbers, whereas for Algorithm 4, it is sufficient that n be a product of small prime numbers.

If one wants to solve a Toeplitz system of linear equations, then other displacement structures may be used. Define

$$Z_\delta = \begin{pmatrix} 0 & 0 & \cdots & 0 & \delta \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix},$$

and let $\Omega = Z_1$ and $\Lambda = Z_{-1}$. Kailath, Kung, Morf [21] show that every Toeplitz matrix satisfies the displacement equation (1.1) with G having nonzero entries only in its first row and last column, a matrix of rank at most 2. Hence, the displacement

rank of a Toeplitz matrix is at most 2 with respect to Z_1 and Z_{-1}. The following result can be found in [17].

PROPOSITION 3.2. *Let $M \in \mathbf{R}^{n\times n}$ be a matrix satisfying the displacement equation*

$$Z_1 \cdot M - M \cdot Z_{-1} = A \cdot B, \tag{3.6}$$

where $A \in \mathbf{R}^{n\times\alpha}$ and $B \in \mathbf{R}^{\alpha\times n}$. Then $\mathcal{F}\cdot M \cdot D_0^{-1}\cdot \mathcal{F}^*$ is a Cauchy-like matrix

$$\mathcal{D}_1 \cdot (\mathcal{F}\cdot M\cdot D_0^{-1}\cdot\mathcal{F}^*) - (\mathcal{F}\cdot M\cdot D_0^{-1}\cdot\mathcal{F}^*)\cdot\mathcal{D}_{-1} = (\mathcal{F}\cdot A)\cdot(B\cdot D_0^*\cdot\mathcal{F}^*), \tag{3.7}$$

where $\mathcal{F} = \sqrt{\frac{1}{n}}\cdot(e^{\frac{2\pi i}{n}(k-1)(j-1)})_{1\le k,j\le n}$ is the normalized inverse discrete Fourier transform matrix

$$\mathcal{D}_1 = \operatorname{diag}\left(1, e^{\frac{2\pi i}{n}}, \ldots, e^{\frac{2\pi i}{n}(n-1)}\right), \quad \mathcal{D}_{-1} = \operatorname{diag}\left(e^{\frac{\pi i}{n}}, e^{\frac{3\pi i}{n}}, \ldots, e^{\frac{(2n-1)\pi i}{n}}\right).$$

Heinig [17] suggests that for a Toeplitz matrix T, one can convert it into the Cauchy-like matrix in (3.7), and Gohberg, Kailath, and Olshevsky [11] suggest that one can rapidly factorize this Cauchy-like matrix using Algorithm 1. The resulting algorithm is called Algorithm GKO in [11]. Since the cost of a fast algorithm for factorizing a Cauchy-like matrix depends linearly on the displacement rank (see Remarks 1 and 2), this method is more efficient than Algorithm 4 if T is given to be a complex matrix.

However, the situation is different if T is real (as often happens in practice). The total cost of complex forward and backward substitution is about $8n^2$ flops; and the total cost of factorizing the Cauchy-like matrix in (3.7) is about $44n^2$ flops for Algorithm 1 (see Remark 2). Using the above procedure, a Toeplitz system is thus solved in about $52n^2$ flops. On the other hand, by treating a Toeplitz matrix as a Toeplitz-plus-Hankel matrix, we can solve a Toeplitz system using Algorithm 4, which completely avoids complex arithmetic. As noted in section 3.1, the cost of Algorithm 4 is about $20.5n^2$ flops for large K, less than half the cost of Algorithm GKO. Furthermore, operating in real arithmetic reduces the storage requirements by half, a big saving for large matrices.

On the other hand, Algorithm GKO does have an advantage over Algorithm 4: it can be shown that the parameter ρ defined in (2.4) is $O(n)$ for (3.7), thus Algorithm GKO could be more accurate. We will address this issue in section 4.

4. Numerical experiments. We have implemented Algorithm 4 in Fortran and have performed a large number of numerical experiments with it to investigate its behavior in finite arithmetic and to compare it with other available algorithms. In this section we discuss some implementation issues and report some of these numerical experiments. We chose Algorithm 3 with $K = 10$ in step 2 of Algorithm 4.

4.1. Implementation issues. A natural way to implement Algorithm 3 is to keep permutations P and Q in vectors and keep both L and U in a single matrix W by storing L in the strict lower triangular part of W (excluding the diagonal) and U upper triangular part (including the diagonal).

However, arrays are stored columnwise in Fortran. Note that U is generated row-by-row in Algorithm 3. In order to store U, columns of W have to be moved

TABLE 4.1
Execution times.

Matrix type	Order n	Execution Time (seconds)				
		GEPP-I	GEPP-II	LEVIN	NEW-I	NEW-II
Type 1	160	$.3\times10^{-1}$	$.3\times10^{-1}$	$.1\times10^{0}$	$.3\times10^{-1}$	$.3\times10^{-1}$
	320	$.2\times10^{0}$	$.2\times10^{0}$	$.4\times10^{0}$	$.9\times10^{-1}$	$.1\times10^{0}$
	640	$.2\times10^{1}$	$.2\times10^{1}$	$.2\times10^{1}$	$.5\times10^{0}$	$.4\times10^{0}$
	1280	$.2\times10^{2}$	$.1\times10^{2}$	$.7\times10^{1}$	$.2\times10^{1}$	$.1\times10^{1}$
	2560	$.1\times10^{3}$	$.7\times10^{2}$	$.3\times10^{2}$	$.1\times10^{2}$	$.6\times10^{1}$
Type 2	160	$.3\times10^{-1}$	$.3\times10^{-1}$	$.1\times10^{0}$	$.2\times10^{-1}$	$.3\times10^{-1}$
	320	$.2\times10^{0}$	$.1\times10^{0}$	$.5\times10^{0}$	$.9\times10^{-1}$	$.9\times10^{-1}$
	640	$.2\times10^{1}$	$.1\times10^{1}$	$.2\times10^{1}$	$.5\times10^{0}$	$.4\times10^{0}$
	1280	$.2\times10^{2}$	$.1\times10^{2}$	$.7\times10^{1}$	$.2\times10^{1}$	$.1\times10^{1}$
	2560	$.1\times10^{3}$	$.7\times10^{2}$	$.3\times10^{2}$	$.1\times10^{2}$	$.6\times10^{1}$
Type 3	160	$.3\times10^{-1}$	$.3\times10^{-1}$	$.1\times10^{0}$	$.3\times10^{-1}$	$.3\times10^{-1}$
	320	$.2\times10^{0}$	$.2\times10^{0}$	$.5\times10^{0}$	$.9\times10^{-1}$	$.9\times10^{-1}$
	640	$.1\times10^{1}$	$.1\times10^{1}$	$.2\times10^{1}$	$.5\times10^{0}$	$.3\times10^{0}$
	1280	$.5\times10^{1}$	$.1\times10^{2}$	$.7\times10^{1}$	$.2\times10^{1}$	$.1\times10^{1}$
	2560	$.3\times10^{2}$	$.7\times10^{2}$	$.3\times10^{2}$	$.1\times10^{2}$	$.5\times10^{1}$
Type 4	160	$.2\times10^{-1}$	$.2\times10^{-1}$	$.1\times10^{0}$	$.3\times10^{-1}$	$.3\times10^{-1}$
	320	$.1\times10^{0}$	$.1\times10^{0}$	$.4\times10^{0}$	$.9\times10^{-1}$	$.1\times10^{0}$
	640	$.1\times10^{1}$	$.1\times10^{1}$	$.2\times10^{1}$	$.5\times10^{0}$	$.3\times10^{0}$
	1280	$.1\times10^{2}$	$.8\times10^{1}$	$.7\times10^{1}$	$.2\times10^{1}$	$.1\times10^{1}$
	2560	*	*	*	$.1\times10^{2}$	$.6\times10^{1}$

into and brought out of fast memory for most steps of elimination for large n. This causes a significant amount of memory traffic between slow and fast memory levels in the memory hierarchy. For more detailed discussions on memory traffic, see, for example, [9, section 2.6].

We reduce this memory traffic by storing rows of U columnwise in Algorithm 3. Let $S \in \mathbf{R}^{n\times n}$ be the matrix that is 1 on the main antidiagonal and 0 everywhere else. For $n = 2$

$$S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

It follows that $\tilde{U} \equiv S \cdot U^T \cdot S$ is an upper triangular matrix, whose kth column is the $(n-k+1)$st row of U in the reverse order. The backward substitution procedure for computing $U^{-1} \cdot y$ in Algorithm 3 can be rewritten as a forward substitution as $S \cdot ((\tilde{U}^T)^{-1} \cdot (S \cdot y))$. Our numerical experiments indicate that this technique speeds up both Algorithm 3 and Algorithm 4 by up to a factor of 2 (see Table 4.1).

Our numerical experiments indicate that Algorithm 2 is slightly less accurate than straightforward GEPP in many cases. Hence, we perform one step of iterative refinement for Algorithm 4 to get the following (see [15, section 3.5]).

ALGORITHM 5. Solving $M \cdot x = z$ with iterative refinement.

1. Compute the factorization $M = \mathcal{Q}_{1,1}^T \cdot P \cdot L \cdot U \cdot Q \cdot \mathcal{Q}_{1,-1}^T$ and the solution $x^{(1)} = \mathcal{Q}_{1,-1} \cdot Q^T \cdot U^{-1} \cdot L^{-1} \cdot P^T \cdot \mathcal{Q}_{1,1} \cdot z$ using Algorithm 4.
2. Compute the residual $r^{(1)} = z - M \cdot x^{(1)}$.
3. Compute the refined solution $x^{(2)} = x^{(1)} + \mathcal{Q}_{1,-1} \cdot Q^T \cdot U^{-1} \cdot L^{-1} \cdot P^T \cdot \mathcal{Q}_{1,1} \cdot r^{(1)}$.

4. Compute $k_{\max} = \text{argmin}_{1 \le k \le 2} \|z - M \cdot x^{(k)}\|$.
5. Return $x^{(k_{\max})}$ as the computed solution.

The norm in step 4 can be any operator norm.

Our numerical experiments show that Algorithm 5 is in general more accurate than Algorithm 4. Since the residual vectors $z - M \cdot x^{(k)}$ can be computed in $O(n \log_2 n)$ flops using convolution (see [23]), the extra cost for computing $x^{(k_{\max})}$ involves basically a forward and a backward substitution, about $2n^2$ flops, an increase of about 10% over the cost of Algorithm 4 (see section 3.1).

4.2. Numerical results. The computations were done on an IBM RS6000 workstation in double precision where the machine precision is $\epsilon \approx 1.1 \times 10^{-16}$.

We compared the following algorithms[3].

- GEPP-I: LAPACK [1] subroutines DGETRF+DGETRS for solving a general dense linear system of equations using GEPP, with Fortran BLAS and without iterative refinement; cost: $O(n^3)$ flops.
- GEPP-II: LAPACK routines DGETRF+DGETRS for solving a general dense linear system of equations using GEPP, with optimized BLAS and one step of iterative refinement; cost: $O(n^3)$ flops.
- LEVIN: The algorithm available on `Netlib`; cost: $O(n^2)$ flops.
- NEW-I: Implementation of Algorithm 4 by storing rows of U row-wise and with no iterative refinement; cost: $O(n^2)$ flops.
- NEW-II: Implementation of Algorithm 4 by storing rows of U columnwise and with one step of iterative refinement; cost: $O(n^2)$ flops.

We solve Toeplitz linear systems of equations $T \cdot x = z$ for random right-hand side vectors z and the following types of Toeplitz matrices $T = (t_{k-j})_{1 \le k,j \le n}$:

- *Type* 1: $\{t_k\}$ randomly generated from uniform distribution on $(0, 1)$. A Type 1 matrix is usually well conditioned.
- *Type* 2: $t_0 = 2\omega$ and $t_k = \frac{\sin(2\pi\omega k)}{\pi k}$ for $k \neq 0$. $\omega \in [0, 1/2]$ is a parameter. A Type 2 matrix is also called the Prolate matrix in [11, 27]; it is very ill conditioned for small ω. In our experiments we took $\omega = 0.25$.
- *Type* 3: $t_k = a^{k^2}$ with $0 < a < 1$. A Type 3 matrix is also called the Gauss matrix in [11]; it is very ill conditioned for a close to 1. In our experiments we took $a = 0.95$.
- *Type* 4: t_0 is randomly generated from uniform distribution in $(0.9, 1)$; $t_k = -t_0$ for $k > 0$; $t_k = 0$ for $-n/2 < k < 0$; and the rest are randomly generated from uniform distribution in $(0, 1)$. The straightforward GEPP produces huge element growth on a Type 4 matrix.

Our numerical results are summarized in Tables 4.1 and 4.2. NEW-II is faster than NEW-I by a factor of up to 2 for large n and for all four types of matrices, and is more accurate than NEW-I for Types 1 and 4 matrices. On the other hand, GEPP-II is as accurate as GEPP-I but is up to a factor of 2 faster. For $n = 2560$, NEW-II is up to 17 times faster than GEPP-I and up to 10 times faster than GEPP-II, respectively, whereas LEVIN is only up to 3 times faster than GEPP-I and up to 2 times faster than GEPP-II, respectively. Both NEW-I and NEW-II solve all linear systems successfully, whereas GEPP-I, GEPP-II, and LEVIN fail on Type 4 matrices.

[3]This list of algorithms does not include those developed by Chandrasekaran and Sayed [5], Gohberg, Kailath, and Olshevsky [11], and Heinig [17]. The algorithm of [5] was implemented in `matlab`, the algorithm of [11] was implemented in C but was inaccessible to us, and the algorithm of [17] was never efficiently implemented.

TABLE 4.2
Relative residuals.

Matrix Type	Order n	$\frac{\|Tx-b\|_1}{\sqrt{n}\cdot\epsilon\cdot(\|T\|_1\cdot\|x\|_1+\|b\|_1)}$				
		GEPP-I	GEPP-II	LEVIN	NEW-I	NEW-II
Type 1	160	$.8\times10^{-1}$	$.9\times10^{-1}$	$.4\times10^{1}$	$.1\times10^{2}$	$.9\times10^{-1}$
	320	$.9\times10^{-1}$	$.8\times10^{-1}$	$.3\times10^{2}$	$.4\times10^{2}$	$.1\times10^{0}$
	640	$.8\times10^{-1}$	$.1\times10^{0}$	$.1\times10^{3}$	$.4\times10^{2}$	$.5\times10^{-1}$
	1280	$.2\times10^{0}$	$.2\times10^{0}$	$.2\times10^{2}$	$.7\times10^{2}$	$.2\times10^{0}$
	2560	$.9\times10^{-1}$	$.9\times10^{-1}$	$.3\times10^{2}$	$.2\times10^{3}$	$.9\times10^{-1}$
Type 2	160	$.8\times10^{-1}$	$.1\times10^{0}$	$.1\times10^{1}$	$.5\times10^{0}$	$.5\times10^{0}$
	320	$.9\times10^{-1}$	$.9\times10^{-1}$	$.1\times10^{1}$	$.4\times10^{0}$	$.4\times10^{0}$
	640	$.8\times10^{-1}$	$.1\times10^{0}$	$.7\times10^{0}$	$.2\times10^{0}$	$.2\times10^{0}$
	1280	$.7\times10^{-1}$	$.9\times10^{-1}$	$.2\times10^{1}$	$.2\times10^{0}$	$.2\times10^{0}$
	2560	$.7\times10^{-1}$	$.7\times10^{-1}$	$.7\times10^{3}$	$.7\times10^{0}$	$.7\times10^{0}$
Type 3	160	$.2\times10^{-1}$	$.3\times10^{-1}$	$.8\times10^{-1}$	$.1\times10^{1}$	$.1\times10^{1}$
	320	$.2\times10^{-1}$	$.2\times10^{-1}$	$.7\times10^{-1}$	$.9\times10^{0}$	$.9\times10^{0}$
	640	$.1\times10^{-1}$	$.2\times10^{-1}$	$.6\times10^{-1}$	$.1\times10^{1}$	$.1\times10^{1}$
	1280	$.9\times10^{-2}$	$.1\times10^{-1}$	$.5\times10^{-1}$	$.5\times10^{0}$	$.5\times10^{0}$
	2560	$.7\times10^{-2}$	$.9\times10^{-2}$	$.3\times10^{-1}$	$.5\times10^{0}$	$.5\times10^{0}$
Type 4	160	$.4\times10^{15}$	$.4\times10^{15}$	$.6\times10^{13}$	$.3\times10^{1}$	$.1\times10^{0}$
	320	$.2\times10^{15}$	$.2\times10^{15}$	$.4\times10^{13}$	$.5\times10^{1}$	$.2\times10^{-1}$
	640	$.2\times10^{15}$	$.2\times10^{15}$	$.2\times10^{13}$	$.1\times10^{2}$	$.4\times10^{-1}$
	1280	$.1\times10^{15}$	$.1\times10^{15}$	$.9\times10^{13}$	$.5\times10^{1}$	$.1\times10^{0}$
	2560	*	*	*	$.4\times10^{1}$	$.2\times10^{-1}$

5. Error analysis. In this section, we do a backward error analysis for Algorithms 1 through 3 by establishing an ∞-norm upper bound on the matrix H in the equation

$$\hat{L}\cdot\hat{U} = C + H, \tag{5.1}$$

where C is the Cauchy-like matrix to be factored, $\hat{L}\cdot\hat{U}$ is the computed LU factorization, and we assume that no pivoting is done. In the following, we first establish some notation and then analyze error propagation by using induction. At the end of this section we will briefly discuss error propagation for Algorithm 4.

5.1. Notation. At the kth step of elimination in finite arithmetic, let $\hat{C}^{(k)} = \begin{pmatrix} \gamma_k & (u^{(k)})^T \\ r^{(k)} & \tilde{C}_k \end{pmatrix}$ be the Cauchy-like matrix satisfying the displacement equation

$$\Omega_k\cdot\hat{C}^{(k)} - \hat{C}^{(k)}\cdot\Lambda_k = \hat{A}^{(k)}\cdot\hat{B}^{(k)} \tag{5.2}$$

with $\Omega_k = \mathrm{diag}(\omega_k, \Omega_{k+1})$ and $\Lambda_k = \mathrm{diag}(\lambda_k, \Lambda_{k+1}) \in \mathbf{R}^{(n-k+1)\times(n-k+1)}$ diagonal;

$$\hat{A}^{(k)} = \begin{pmatrix} (\hat{a}_k^{(k)})^T \\ \tilde{A}_{k+1} \end{pmatrix} = \begin{pmatrix} (\hat{a}_k^{(k)})^T \\ \vdots \\ (\hat{a}_n^{(k)})^T \end{pmatrix}, \quad \hat{B}^{(k)} = \begin{pmatrix} \hat{b}_k^{(k)} & \tilde{B}_{k+1} \end{pmatrix} = \begin{pmatrix} \hat{b}_k^{(k)}, \cdots, \hat{b}_n^{(k)} \end{pmatrix}.$$

For $k = 1$ we drop the superscripts so that $\hat{C}^{(1)} = C$, $\hat{A}^{(1)} = A$, $\hat{B}^{(1)} = B$, etc., and (5.2) reduces to (2.2).

To perform elimination we write

$$\hat{C}^{(k)} = \begin{pmatrix} 1 & 0 \\ l^{(k)} & I \end{pmatrix} \cdot \begin{pmatrix} \gamma_k & (u^{(k)})^T \\ 0 & C^{(k+1)} \end{pmatrix},$$

where $l^{(k)} = r^{(k)}/\gamma_k$ and $C^{(k+1)} = \tilde{C}_{k+1} - l^{(k)} \cdot (u^{(k)})^T$ satisfies the displacement equation

$$\Omega_{k+1} \cdot C^{(k+1)} - C^{(k+1)} \cdot \Lambda_{k+1} = A^{(k+1)} \cdot B^{(k+1)}$$

with $A^{(k+1)} = \tilde{A}_{k+1} - l^{(k)} \cdot (\hat{a}_k^{(k)})^T$ and $B^{(k+1)} = \tilde{B}_{k+1} - \hat{b}_k^{(k)} \cdot (u^{(k)})^T/\gamma_k$.

Let the computed γ_k, $r^{(k)}$, and $u^{(k)}$ be $\hat{\gamma}_k$, $\hat{r}^{(k)}$, and $\hat{u}^{(k)}$, and let $\hat{l}^{(k)} = \mathbf{fl}(\hat{r}^{(k)}/\hat{\gamma}_k)$. For $k = 1$, we write $r = r^{(1)}$, $u = u^{(1)}$, $l = l^{(1)}$, and $\hat{r} = \hat{r}^{(1)}$, $\hat{u} = \hat{u}^{(1)}$, $\hat{l} = \hat{l}^{(1)}$.

Let $\bar{G}^{(k+1)} = \bar{A}^{(k+1)} \cdot \bar{B}^{(k+1)}$ with $\bar{A}^{(k+1)} = \tilde{A}_{k+1} - \hat{l}^{(k)} \cdot (\hat{a}_k^{(k)})^T$ and $\bar{B}^{(k+1)} = \tilde{B}_{k+1} - \hat{b}_k^{(k)} \cdot (\hat{u}^{(k)})^T/\hat{\gamma}_k$. The generator at the $(k+1)$st step is $\hat{G}^{(k+1)} = \hat{A}^{(k+1)} \cdot \hat{B}^{(k+1)}$. For Algorithm 1, $\hat{A}^{(k+1)} = \mathbf{fl}\left(\bar{A}^{(k+1)}\right)$ and $\hat{B}^{(k+1)} = \mathbf{fl}\left(\bar{B}^{(k+1)}\right)$, and for Algorithm 3, $\hat{A}^{(k+1)}$ is the computed Q factor in the QR factorization of $\mathbf{fl}(\bar{A}^{(k+1)})$ and $\hat{B}^{(k+1)}$ is the product of the R factor and $\mathbf{fl}(\bar{B}^{(k+1)})$. We further define $\bar{C}^{(k+1)}$ to be the matrix satisfying the displacement equation

$$\Omega_{k+1} \cdot \bar{C}^{(k+1)} - \bar{C}^{(k+1)} \cdot \Lambda_{k+1} = \bar{G}^{(k+1)}. \tag{5.3}$$

Define

$$\tau = \max_{2 \le k \le n} \|\hat{G}^{(k)} - \bar{G}^{(k)}\|_\infty, \qquad \mu = \max_{2 \le k \le n} \|\hat{C}^{(k)}\|_{\max}; \tag{5.4}$$

and

$$\nu = \max_{1 \le k \le i \le n} \frac{|\hat{a}_i^{(k)}|^T \cdot |\hat{b}_k^{(k)}|}{|\hat{\gamma}_k|}, \qquad \psi = \frac{\max_{1 \le k \le n} \left\| |\hat{A}^{(k)}| \cdot |\hat{B}^{(k)}| \right\|_{\max}}{\max_{1 \le k \le n} \left\| \hat{A}^{(k)} \cdot \hat{B}^{(k)} \right\|_{\max}}. \tag{5.5}$$

τ is a measure of the accuracy in computing the generators; μ is a measure of element growth in the LU factorization, since it is easy to show that

$$\mu \le \| |\hat{L}| \cdot |\hat{U}| \|_{\max} + O(\epsilon); \tag{5.6}$$

$\psi \ge 1$ is of order 1 in general, but it could happen that $\psi \gg 1$ if both $\hat{A}^{(k)}$ and $\hat{B}^{(k)}$ are ill conditioned for some k. We will further discuss these four parameters in sections 5.4 and 5.5.

LEMMA 5.1. *For any* $1 \le k \le i, j \le n$

$$|\hat{a}_i^{(k)}|^T \cdot |\hat{b}_j^{(k)}| \le \xi_{\max} \cdot \psi \cdot \mu.$$

Proof. Let $|(\hat{a}_s^{(f)})^T \cdot \hat{b}_m^{(f)}| = \max_{1 \le h \le n} \|\hat{A}^{(h)} \cdot \hat{B}^{(h)}\|_{\max}$. Then

$$\begin{aligned} |\hat{a}_i^{(k)}|^T \cdot |\hat{b}_j^{(k)}| &\le \psi \cdot |(a_s^{(f)})^T \cdot b_m^{(f)}| \le \psi \cdot \|C^{(f)}\|_{\max} \cdot |\omega_s - \lambda_m| \\ &\le \xi_{\max} \cdot \psi \cdot \mu. \qquad \square \end{aligned}$$

5.2. Error propagation for one step of elimination. Let $\hat{L}^{(2)} \cdot \hat{U}^{(2)} = \hat{C}^{(2)} + H^{(2)}$ be the computed LU factorization of $\hat{C}^{(2)}$. Then the computed LU factorization of C satisfies

$$\hat{L} = \begin{pmatrix} 1 & 0 \\ \hat{l} & \hat{L}^{(2)} \end{pmatrix} \text{ and } \hat{U} = \begin{pmatrix} \hat{\gamma}_1 & \hat{u}^T \\ 0 & \hat{U}^{(2)} \end{pmatrix}.$$

It follows that

$$\begin{aligned} \hat{L} \cdot \hat{U} &= \begin{pmatrix} \hat{\gamma}_1 & \hat{u}^T \\ \hat{l}\hat{\gamma}_1 & \hat{L}^{(2)} \cdot \hat{U}^{(2)} + \hat{l} \cdot \hat{u}^T \end{pmatrix} \\ &= \begin{pmatrix} \gamma_1 & u^T \\ r & \tilde{C}_2 \end{pmatrix} + \begin{pmatrix} \hat{\gamma}_1 - \gamma_1 & (\hat{u} - u)^T \\ \hat{l}\hat{\gamma}_1 - r & H^{(2)} + (\hat{C}^{(2)} - \bar{C}^{(2)}) + (\bar{C}^{(2)} - \tilde{C}_2 + \hat{l} \cdot \hat{u}^T) \end{pmatrix}. \end{aligned}$$

Since $M = \begin{pmatrix} \gamma_1 & u^T \\ r & \tilde{C}_2 \end{pmatrix}$, (5.1) and the last equation imply

$$H = \begin{pmatrix} \hat{\gamma}_1 - \gamma_1 & (\hat{u} - u)^T \\ \hat{l}\hat{\gamma}_1 - r & H^{(2)} + (\hat{C}^{(2)} - \bar{C}^{(2)}) + (\bar{C}^{(2)} - \tilde{C}_2 + \hat{l} \cdot \hat{u}^T) \end{pmatrix}. \tag{5.7}$$

For the rest of section 5.2, we bound $|\hat{\gamma}_1 - \gamma_1|$, $|\hat{u} - u|$, $|\hat{l}\hat{\gamma}_1 - r|$, $|\hat{C}^{(2)} - \bar{C}^{(2)}|$, and $|\bar{C}^{(2)} - \tilde{C}_2 + \hat{l} \cdot \hat{u}^T|$. We obtain an upper bound on H by induction in section 5.3.

Displacement equation (5.2) for $k = 1$ implies that

$$\gamma_1 = (\omega_1 - \lambda_1)^{-1} \cdot a_1^T \cdot b_1,\ r = (\Omega_1 - \lambda_1 I)^{-1} \cdot \tilde{A}_2 \cdot b_1,\ u = (\omega_1 I - \Lambda_1)^{-1} \cdot \tilde{B}_2^T \cdot a_1. \tag{5.8}$$

For both Algorithm 1 and Algorithm 3, the errors in these quantities and l can be bounded as follows, using our model of arithmetic (1.3) and Lemma 5.1:

$$\begin{aligned} |\hat{\gamma}_1 - \gamma_1| &\le \alpha\eta|\omega_1 - \lambda_1|^{-1} \cdot |a_1|^T \cdot |b_1| \le \frac{\alpha\eta\psi\mu\xi_{\max}}{\xi_{\min}}, \\ |\hat{r} - r| &\le \alpha\eta|\Omega_1 - \lambda_1 I|^{-1} \cdot |\tilde{A}_2| \cdot |b_1| \le \frac{\alpha\eta\psi\mu\xi_{\max}}{\xi_{\min}} \cdot e, \\ |\hat{u} - u| &\le \alpha\eta|\omega_1 I - \Lambda_1|^{-1} \cdot |\tilde{B}_2^T| \cdot |a_1| \le \frac{\alpha\eta\psi\mu\xi_{\max}}{\xi_{\min}} \cdot e, \end{aligned} \tag{5.9}$$

and $|\hat{l} - \hat{r}/\hat{\gamma}_1| \le \eta|\hat{l}|$, where[4] η is a small multiple of ϵ, and $e = (1, \ldots, 1)^T$. These relations imply that

$$\begin{aligned} |\hat{l}\hat{\gamma}_1 - r| &\le |\hat{l}\hat{\gamma}_1 - \hat{r}| + |\hat{r} - r| \\ &\le \eta|\hat{\gamma}_1| \cdot |\hat{l}| + \frac{\alpha\eta\psi\mu\xi_{\max}}{\xi_{\min}} \cdot e. \end{aligned} \tag{5.10}$$

Since

$$\Omega_2 \cdot \bar{C}^{(2)} - \bar{C}^{(2)} \cdot \Lambda_2 = \bar{G}^{(2)} \text{ and } \Omega_2 \cdot \hat{C}^{(2)} - \hat{C}^{(2)} \cdot \Lambda_2 = \hat{G}^{(2)},$$

[4]Throughout section 5 we use the same η in several similar error bounds; hence, η is in fact the maximum of all these different ηs.

by subtracting these two equations we get

$$\Omega_2 \cdot \left(\hat{C}^{(2)} - \bar{C}^{(2)}\right) - \left(\hat{C}^{(2)} - \bar{C}^{(2)}\right) \cdot \Lambda_2 = \hat{G}^{(2)} - \bar{G}^{(2)}.$$

By definition (5.4) this implies

$$\|\hat{C}^{(2)} - \bar{C}^{(2)}\|_\infty \leq \frac{\tau}{\xi_{\min}}. \tag{5.11}$$

Finally, we bound $\bar{H}^{(2)} \equiv \bar{C}^{(2)} - \tilde{C}_2 + \hat{l} \cdot \hat{u}^T$. Write $\hat{l} = (\hat{l}_2, \ldots, \hat{l}_n)^T$ and $\hat{u} = (\hat{u}_2, \ldots, \hat{u}_n)^T$. According to (5.2) and (5.3), the $(i-1, j-1)$ entries of matrices $\bar{C}^{(2)}$, $\tilde{C}$, and $\hat{l} \cdot \hat{u}^T$ are $(a_i^T - \hat{l}_i \cdot a_1^T) \cdot (b_j - b_1 \cdot \hat{u}_j/\hat{\gamma}_1)/(\omega_i - \lambda_j)$, $a_i^T \cdot b_j/(\omega_i - \lambda_j)$, and $\hat{l}_i \cdot \hat{u}_j$ for $2 \leq i, j \leq n$. Thus

$$\begin{aligned} |\bar{H}^{(2)}_{i-1,j-1}| &= \left| \frac{(a_i^T - \hat{l}_i \cdot a_1^T) \cdot (b_j - b_1 \cdot \hat{u}_j/\hat{\gamma}_1) - a_i^T \cdot b_j}{\omega_i - \lambda_j} + \hat{l}_i \cdot \hat{u}_j \right| \\ &= \left| \frac{\hat{l}_i \hat{u}_j(\omega_i - \lambda_j + a_1^T \cdot b_1/\hat{\gamma}_1) - \hat{l}_i a_1^T \cdot b_j - \hat{u}_j \cdot a_i^T b_1/\hat{\gamma}_1}{\omega_i - \lambda_j} \right|. \end{aligned}$$

Equation (5.8) implies that

$$a_1^T \cdot b_1 = (\omega_1 - \lambda_1) \cdot \gamma_1,\ a_1^T \cdot b_j = (\omega_1 - \lambda_j) \cdot u_j, \text{ and } a_i^T \cdot b_1 = (\omega_i - \lambda_1) \cdot r_i.$$

Plugging these relations into the above and simplifying we have

$$\begin{aligned} |\bar{H}^{(2)}_{i-1,j-1}| &= \left| \frac{\hat{l}_i \hat{u}_j \cdot \frac{(\omega_1 - \lambda_1)(\gamma_1 - \hat{\gamma}_1)}{\hat{\gamma}_1} - \hat{l}_i(\omega_1 - \lambda_j)(u_j - \hat{u}_j) - \hat{u}_j \cdot \frac{(\omega_i - \lambda_1)(r_i - \hat{l}_i\hat{\gamma}_1)}{\hat{\gamma}_1}}{\omega_i - \lambda_j} \right| \\ &\leq \frac{|\hat{l}_i| \cdot |\hat{u}_j| \cdot \frac{|\omega_1 - \lambda_1| \cdot |\gamma_1 - \hat{\gamma}_1|}{|\hat{\gamma}_1|} + |\hat{l}_i| \cdot |u_j - \hat{u}_j| \cdot |\omega_1 - \lambda_j| + |\hat{u}_j| \cdot \frac{|\omega_i - \lambda_1| \cdot |r_i - \hat{l}_i\hat{\gamma}_1|}{|\hat{\gamma}_1|}}{\xi_{\min}}. \end{aligned}$$

By relation (5.9) and definition (5.5) we have

$$\frac{|\omega_1 - \lambda_1| \cdot |\gamma_1 - \hat{\gamma}_1|}{|\hat{\gamma}_1|} \leq \frac{\alpha\eta |a_1|^T \cdot |b_1|}{|\hat{\gamma}_1|} \leq \alpha\eta\nu;$$

further, by using relation (5.9) we have

$$\frac{|\omega_i - \lambda_1| \cdot |r_i - \hat{\gamma}_1\hat{l}_i|}{|\hat{\gamma}_1|} \leq \eta|\omega_i - \lambda_1| \cdot |\hat{l}_i| + \frac{\alpha\eta |a_i|^T \cdot |b_1|}{|\hat{\gamma}_1|} \leq \eta\xi_{\max} \cdot |\hat{l}_i| + \alpha\eta\nu.$$

Plugging these relations into the last bound on $|\bar{H}^{(2)}_{i-1,j-1}|$, rewriting the result in matrix form, and simplifying we have

$$\begin{aligned} |\bar{H}^{(2)}| &\leq \frac{\alpha\eta\nu \cdot |\hat{l}| \cdot |\hat{u}|^T + |\hat{l}| \cdot |u - \hat{u}|^T \cdot |\omega_1 I - \Lambda_2| + \eta\xi_{\max}|\hat{l}| \cdot |\hat{u}|^T + \alpha\eta\nu \cdot e \cdot |\hat{u}|^T}{\xi_{\min}} \\ &\leq \frac{(\alpha\nu + \xi_{\max})\eta}{\xi_{\min}} \cdot |\hat{l}| \cdot |\hat{u}|^T + \frac{\alpha\eta}{\xi_{\min}} \cdot |\hat{l}| \cdot |a_1|^T \cdot |\tilde{B}_2| + \frac{\alpha\nu\eta}{\xi_{\min}} \cdot e \cdot |\hat{u}|^T, \end{aligned}$$

where we have used relations (5.9). According to Lemma 5.1, the last relation can be further simplified to

$$|\bar{H}^{(2)}| \le \frac{(\alpha\nu + \xi_{\max})\eta}{\xi_{\min}} \cdot |\hat{l}| \cdot |\hat{u}|^T + \frac{\alpha\eta\psi\mu\xi_{\max}}{\xi_{\min}} \cdot |\hat{l}| \cdot e^T + \frac{\alpha\nu\eta}{\xi_{\min}} \cdot e \cdot |\hat{u}|^T$$

$$\le \frac{\alpha\eta(\nu + \psi \cdot \xi_{\max})}{\xi_{\min}} \cdot (|\hat{l}| + e) \cdot (|\hat{u}| + \mu \cdot e)^T. \tag{5.12}$$

5.3. Error analysis for factorizing Cauchy-like matrices. Let $X_n \in \mathbf{R}^{n\times n}$ and $Y_n \in \mathbf{R}^{n\times n}$ be lower and upper triangular matrices such that

$$X_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 1 & \cdots & 1 & 1 \end{pmatrix} \quad \text{and} \quad Y_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

We also define

$$\Delta^{(k)} = \sum_{s=k+1}^{n} \begin{pmatrix} 0 & 0 \\ 0 & |\bar{C}^{(s)} - \hat{C}^{(s)}| \end{pmatrix}.$$

The following theorem gives an upper bound on $\|H\|\infty$ in (5.1).

THEOREM 5.2. *The backward error H in the LU factorization of a Cauchy-like matrix C in* (5.1) *satisfies*

$$\|H\|_\infty \le \frac{\alpha\eta(\nu + \psi \cdot \xi_{\max})}{\xi_{\min}} \cdot \left\| \left(|\hat{L}| + X_n\right) \cdot \left(|\hat{U}| + \mu \cdot Y_n\right) \right\|_\infty + \frac{n \cdot \tau}{\xi_{\min}}, \tag{5.13}$$

where ν, ψ, μ, and τ are defined in (5.4) *and* (5.5).

Proof. We shall first show that

$$|H| \le \frac{\alpha\eta(\nu + \psi \cdot \xi_{\max})}{\xi_{\min}} \cdot \left(|\hat{L}| + X_n\right) \cdot \left(|\hat{U}| + \mu \cdot Y_n\right) + \frac{\Delta^{(1)}}{\xi_{\min}} \tag{5.14}$$

by using induction on n. We shall then prove the theorem by taking ∞-norm on both sides of (5.14).

Relation (5.14) clearly holds for all Cauchy-like matrices of dimension $n = 1, 2$, and we assume it holds for $n - 1$ as well. In light of (5.7) we have

$$|H| \le \begin{pmatrix} |\hat{\gamma}_1 - \gamma_1| & |\hat{u} - u|^T \\ |\hat{l}\hat{\gamma}_1 - r| & |H^{(2)}| + |\hat{C}^{(2)} - \bar{C}^{(2)}| + |\bar{C}^{(2)} - \tilde{C}_2 + \hat{l} \cdot \hat{u}^T| \end{pmatrix}.$$

Plugging relations (5.9) through (5.12) into the above we have

$$\begin{aligned} |H| &\le \begin{pmatrix} 0 & 0 \\ 0 & |H^{(2)}| \end{pmatrix} + \begin{pmatrix} |\hat{\gamma}_1 - \gamma_1| & |\hat{u} - u|^T \\ |\hat{l}\hat{\gamma}_1 - r| & 0 \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 & 0 \\ 0 & |\hat{C}^{(2)} - \bar{C}^{(2)}| \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & |\bar{C}^{(2)} - \tilde{C}_2 + \hat{l} \cdot \hat{u}^T| \end{pmatrix} \\ &\le \begin{pmatrix} 0 & 0 \\ 0 & |H^{(2)}| \end{pmatrix} + \frac{\alpha\eta\psi\xi_{\max}}{\xi_{\min}} \cdot \begin{pmatrix} \mu & \mu \cdot e^T \\ \mu \cdot e & 0 \end{pmatrix} + \eta \cdot \begin{pmatrix} 0 & 0 \\ |\hat{l}| \cdot |\hat{\gamma}_1| & 0 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&+\begin{pmatrix} 0 & 0 \\ 0 & |\hat{C}^{(2)} - \bar{C}^{(2)}| \end{pmatrix} + \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \begin{pmatrix} 0 & 0 \\ 0 & (|\hat{l}| + e)\cdot(|\hat{u}| + \mu\cdot e)^T \end{pmatrix} \\
&\le \begin{pmatrix} 0 & 0 \\ 0 & |H^{(2)}| \end{pmatrix} + \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \begin{pmatrix} 2 \\ |\hat{l}| + e \end{pmatrix} \cdot \begin{pmatrix} |\hat{\gamma}_1| + \mu & (|\hat{u}| + \mu\cdot e)^T \end{pmatrix} \\
&+\begin{pmatrix} 0 & 0 \\ 0 & |\hat{C}^{(2)} - \bar{C}^{(2)}| \end{pmatrix}, \qquad (5.15)
\end{aligned}
$$

where we have used the fact that $\xi_{\max} \ge \xi_{\min}$ and $\psi \ge 1$. The induction hypothesis implies that

$$
\|H^{(2)}| \le \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \left(|\hat{L}^{(2)}| + X_{n-1}\right) \cdot \left(|\hat{U}^{(2)}| + \mu\cdot Y_{n-1}\right) + \frac{\Delta^{(2)}}{\xi_{\min}}.
$$

Plugging this relation into (5.15) we have

$$
\begin{aligned}
|H| &\le \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \begin{pmatrix} 0 \\ |\hat{L}^{(2)}| + X_{n-1} \end{pmatrix} \cdot \begin{pmatrix} 0 & |\hat{U}^{(2)}| + \mu\cdot Y_{n-1} \end{pmatrix} \\
&\quad + \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \begin{pmatrix} 2 \\ |\hat{l}| + e \end{pmatrix} \cdot \begin{pmatrix} |\hat{\gamma}_1| + \mu & (|\hat{u}| + \mu\cdot e)^T \end{pmatrix} \\
&\quad + \frac{\Delta^{(1)}}{\xi_{\min}} \\
&= \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \begin{pmatrix} 1+1 & 0 \\ |\hat{l}| + e & |\hat{L}^{(2)}| + X_{n-1} \end{pmatrix} \cdot \begin{pmatrix} |\hat{\gamma}_1| + \mu & (|\hat{u}| + \mu\cdot e)^T \\ 0 & |\hat{U}^{(2)}| + \mu\cdot Y_{n-1} \end{pmatrix} \\
&\quad + \frac{\Delta^{(2)}}{\xi_{\min}} \\
&= \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \left(|\hat{L}| + X_n\right) \cdot (|\hat{U}| + \mu\cdot Y_n) + \frac{\Delta^{(1)}}{\xi_{\min}}.
\end{aligned}
$$

Hence, relation (5.14) holds for all n. Taking ∞-norm on both of its sides,

$$
\begin{aligned}
\|H\|_\infty &\le \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \left\| \left(|\hat{L}| + X_n\right) \cdot \left(|\hat{U}| + \mu\cdot Y_n\right) \right\|_\infty + \frac{\left\|\Delta^{(1)}\right\|_\infty}{\xi_{\min}} \\
&\le \frac{\alpha\eta(\nu + \psi\cdot\xi_{\max})}{\xi_{\min}} \cdot \left\| \left(|\hat{L}| + X_n\right) \cdot \left(|\hat{U}| + \mu\cdot Y_n\right) \right\|_\infty + \frac{n\cdot\tau}{\xi_{\min}}. \qquad \square
\end{aligned}
$$

It follows from (5.6) that μ is an upper bound on the element growth in the computed LU factorization. Thus Theorem 5.2 shows that the backward error in the computed LU factorization is bounded by $(\alpha\eta(\nu + \psi\cdot\xi_{\max}))/\xi_{\min}$ times the element growth in $|\hat{L}|$ and $|\hat{U}|$ plus the error in computing the generators.

5.4. Error analysis for Algorithm 1. In this subsection, we assume that partial pivoting has been done before hand so that Algorithm 1 does not perform any pivoting.

For Algorithm 1, the generators are computed as

$$
\hat{A}^{(k+1)} = \mathbf{fl}(\bar{A}^{(k+1)}) \quad \text{and} \quad \hat{B}^{(k+1)} = \mathbf{fl}(\bar{B}^{(k+1)}),
$$

where $\bar{A}^{(k+1)} = \tilde{A}_{k+1} - \hat{l}^{(k)}\cdot(\hat{a}_k^{(k)})^T$ and $\bar{B}^{(k+1)} = \tilde{B}_{k+1} - \hat{b}_k^{(k)}\cdot(\hat{u}^{(k)})^T/\hat{\gamma}_k$. It follows that

$$
|\hat{A}^{(k+1)} - \bar{A}^{(k+1)}| \le \eta\left(|\tilde{A}_{k+1}| + |\hat{l}^{(k)}|\cdot|\hat{a}_k^{(k)}|^T\right),
$$

$$|\hat{B}^{(k+1)} - \bar{B}^{(k+1)}| \le \eta \left(|\tilde{B}_{k+1}| + |\hat{b}_k^{(k)}| \cdot |\hat{u}^{(k)}|^T / |\hat{\gamma}_k| \right).$$

Hence,

$$\begin{aligned}
|\hat{G}^{(k+1)} - \bar{G}^{(k+1)}| &= |\hat{A}^{(k+1)} \cdot \hat{B}^{(k+1)} - \bar{A}^{(k+1)} \cdot \bar{B}^{(k+1)}| \\
&\le |\hat{A}^{(k+1)} - \bar{A}^{(k+1)}| \cdot |\bar{B}^{(k+1)}| + |\bar{A}^{(k+1)}| \cdot |\hat{B}^{(k+1)} - \bar{B}^{(k+1)}| + O(\epsilon^2) \\
&\le 2\eta \left(|\tilde{A}_{k+1}| + |\hat{l}^{(k)}| \cdot |\hat{a}_k^{(k)}|^T \right) \cdot \left(|\tilde{B}_{k+1}| + |\hat{b}_k^{(k)}| \cdot |\hat{u}^{(k)}|^T / |\hat{\gamma}_k| \right) + O(\epsilon^2) \\
&= 2\eta \left(|\tilde{A}_{k+1}| \cdot |\tilde{B}_{k+1}| + (|\tilde{A}_{k+1}| \cdot |\hat{b}_k^{(k)}| / |\hat{\gamma}_k|) \cdot |\hat{u}^{(k)}|^T \right) \\
&\quad + 2\eta |\hat{l}^{(k)}| \cdot \left(|\hat{a}_k^{(k)}|^T \cdot |\tilde{B}_{k+1}| + (|\hat{a}_k^{(k)}|^T \cdot |\hat{b}_k^{(k)}| / |\hat{\gamma}_k|) \cdot |\hat{u}^{(k)}|^T \right) + O(\epsilon^2).
\end{aligned}$$

Using definitions (5.4) and (5.5), Lemma 5.1, and the fact that $|\hat{l}^{(k)}| \le e + O(\epsilon)$, we have

$$\begin{aligned}
\|\hat{G}^{(k+1)} - \bar{G}^{(k+1)}\|_{\max} &\le 2\eta \left(\| |\tilde{A}_{k+1}| \cdot |\tilde{B}_{k+1}| \|_{\max} + \nu \| e \cdot |\hat{u}^{(k)}|^T \|_{\max} \right) \\
&\quad + 2\eta \left(\| e \cdot |\hat{a}_k^{(k)}|^T \cdot |\tilde{B}_{k+1}| \|_{\max} + \nu \cdot \| e \cdot |\hat{u}^{(k)}|^T \|_{\max} \right) + O(\epsilon^2) \\
&\le 2\eta(\psi\mu\xi_{\max} + \nu\mu + \psi\mu\xi_{\max} + \nu\mu) + O(\epsilon^2) \\
&= 4\eta\mu(\nu + \psi\xi_{\max}) + O(\epsilon^2) \ . \qquad (5.16)
\end{aligned}$$

According to definition (5.4), this implies that for Algorithm 1

$$\tau \le 4n\eta\mu(\nu + \psi\xi_{\max}) + O(\epsilon^2).$$

Plugging this into (5.7), and using the fact that $\| |\hat{L}| + X_n \|_\infty \le 2n + O(\epsilon)$, we have the following.

THEOREM 5.3. *For Algorithm* 1*, the backward error* H *in the* LU *factorization of* C *in* (5.1) *satisfies*

$$\|H\|_\infty \le \frac{2n\eta \cdot (\alpha + 2) \cdot (\nu + \psi \cdot \xi_{\max})}{\xi_{\min}} \cdot \left(\|\hat{U}\|_\infty + n \cdot \mu \right) + O(\epsilon^2).$$

One expects ψ to be of the order 1 in general. The fact that Algorithm 1 performs partial pivoting means that

$$|\hat{\gamma}_k| = \left| \mathbf{fl} \left(\frac{(a_k^{(k)})^T \cdot b_k^{(k)}}{\omega_k - \lambda_k} \right) \right| \ge \left| \mathbf{fl} \left(\frac{(a_i^{(k)})^T \cdot b_k^{(k)}}{\omega_i - \lambda_k} \right) \right|$$

for $1 \le k \le i \le n$. Comparing this with the definition for ν in (5.5), one expects ν to be of the order $\xi_{\max}$ in general. Hence, Theorem 5.3 suggests that in general the backward error for Algorithm 1 is of the order $\epsilon \cdot \xi_{\max}/\xi_{\min} \cdot \|\hat{U}\|_\infty$.

However, if both $\hat{A}^{(k)}$ and $\hat{B}^{(k)}$ are ill conditioned for some k, it could happen that $\psi \gg 1$ and $\nu \gg \xi_{\max}$. If this happens, then the backward error for Algorithm 1 could be much larger.

On the other hand, if the straightforward GEPP is applied to C, then the backward error is basically $\epsilon \cdot \|\hat{U}\|_\infty$. Thus Algorithm 1 appears to be less numerically stable than straightforward GEPP on C. These conclusions are consistent with those of Sweet and Brent [26].

5.5. Error analysis for Algorithm 2. In this subsection we assume that partial pivoting has been done before hand, so that Algorithm 2 does not perform any pivoting.

For Algorithm 2, $\hat{A}^{(k+1)}$ is the computed Q factor in the QR factorization of $\mathbf{fl}(\bar{A}^{(k+1)})$ and $\hat{B}^{(k+1)}$ is the product of the R factor and $\mathbf{fl}(\bar{B}^{(k+1)})$. In finite arithmetic, let $\mathbf{fl}(\bar{A}^{(k+1)}) = \hat{A}^{(k+1)} \cdot R + E_1$ be the QR factorization of $\mathbf{fl}(\bar{A}^{(k+1)})$, and $\hat{B}^{(k+1)} = R \cdot \mathbf{fl}(\bar{B}^{(k+1)}) + E_2$. It is known [15] that the error matrices satisfy

$$\|E_1\|_2 \le \eta_1 \alpha \cdot n \cdot \|\mathbf{fl}(\bar{A}^{(k+1)})\|_2 \quad \text{and} \quad \|E_2\|_\infty \le \eta_2 \alpha \cdot \|R\|_\infty \cdot \|\mathbf{fl}(\bar{B}^{(k+1)})\|_\infty, \tag{5.17}$$

where η_1 and η_2 are small multiples of ϵ. We observe that, after some algebra,

$$\hat{A}^{(k+1)} \cdot \hat{B}^{(k+1)} - \mathbf{fl}(\bar{A}^{(k+1)}) \cdot \mathbf{fl}(\bar{B}^{(k+1)}) = -E_1 \cdot \mathbf{fl}(\bar{B}^{(k+1)}) + \hat{A}^{(k+1)} \cdot E_2. \tag{5.18}$$

In the following, we shall derive an upper bound for τ. To this end, we need to derive norm bounds for some of the related quantities. Since Algorithm 2 performs row pivoting and keeps $\hat{A}^{(k)}$ numerically column orthogonal at every step, we have $\|\hat{l}^{(k)}\|_{\max} \le 1 + O(\epsilon)$,

$$\|\mathbf{fl}(\bar{A}^{(k+1)})\|_\infty \le \|\tilde{A}_{k+1}\|_\infty + \|\hat{l}^{(k)} \cdot (\hat{a}_k^{(k)})^T\|_\infty + O(\epsilon) \le 2\sqrt{\alpha} + O(\epsilon), \tag{5.19}$$

and

$$\|R\|_\infty = \|(\hat{A}^{(k+1)})^T \cdot \left(\tilde{A}_{k+1} - \hat{l}^{(k)} \cdot (\hat{a}_k^{(k)})^T\right)\|_\infty + O(\epsilon) \le \sqrt{\alpha} \cdot (\sqrt{n}+1) + O(\epsilon). \tag{5.20}$$

Since $\hat{A}^{(k)}$ is numerically column orthogonal, it follows that

$$\|(\Omega_k - \lambda_k I)^{-1} \cdot \hat{A}^{(k)} \cdot \hat{b}_k^{(k)}\|_2 \ge \frac{\|\hat{A}^{(k)} \cdot \hat{b}_k^{(k)}\|_2}{\xi_{\max}} = \frac{\|\hat{b}_k^{(k)}\|_2}{\xi_{\max}} + O(\epsilon).$$

The fact that Algorithm 2 performs row pivoting gives

$$\begin{aligned} |\hat{\gamma}_k| &= \|(\Omega_k - \lambda_k I)^{-1} \cdot \hat{A}^{(k)} \cdot \hat{b}_k^{(k)}\|_{\max} + O(\epsilon) \\ &\ge \frac{1}{\sqrt{n}} \cdot \|(\Omega_k - \lambda_k I)^{-1} \cdot \hat{A}^{(k)} \cdot \hat{b}_k^{(k)}\|_2 + O(\epsilon) \\ &\ge \frac{\|\hat{b}_k^{(k)}\|_2}{\sqrt{n} \cdot \xi_{\max}} + O(\epsilon). \end{aligned} \tag{5.21}$$

With these relations we get

$$\begin{aligned} \|\mathbf{fl}(\bar{B}^{(k+1)})\|_\infty &\le \|\tilde{B}_{k+1}\|_\infty + \|\hat{b}_k^{(k)} \cdot (\hat{u}^{(k)})^T\|_\infty / |\hat{\gamma}_k| + O(\epsilon) \\ &\le \sqrt{\alpha} \cdot \|\hat{B}^{(k)}\|_2 + \|\hat{b}_k^{(k)}\|_2 \cdot \|\hat{U}\|_\infty / |\hat{\gamma}_k| + O(\epsilon) \\ &= \sqrt{\alpha} \cdot \|\hat{A}^{(k)} \cdot \hat{B}^{(k)}\|_2 + \|\hat{b}_k^{(k)}\|_2 \cdot \|\hat{U}\|_\infty / |\hat{\gamma}_k| + O(\epsilon) \\ &\le \sqrt{\alpha} \cdot n \|\hat{A}^{(k)} \cdot \hat{B}^{(k)}\|_{\max} + \sqrt{n} \cdot \xi_{\max} \cdot \|\hat{U}\|_\infty + O(\epsilon) \\ &\le \sqrt{\alpha} \cdot n \cdot \xi_{\max} \|\hat{C}^{(k)}\|_{\max} + \sqrt{n} \cdot \xi_{\max} \cdot \|\hat{U}\|_\infty + O(\epsilon) \\ &\le \sqrt{\alpha} \cdot n \cdot \xi_{\max} \cdot \mu + \sqrt{n} \cdot \xi_{\max} \cdot \|\hat{U}\|_\infty + O(\epsilon). \end{aligned} \tag{5.22}$$

To obtain an upper bound on τ, we now take ∞-norm on both sides of (5.18). Using relations (5.17) through (5.20) and (5.22), the right-hand side of (5.18) is bounded

above by

$$\begin{aligned}
&\|E_1\|_\infty \cdot \|\mathbf{fl}(\bar{B}^{(k+1)})\|_\infty + \|\hat{A}^{(k+1)}\|_\infty \cdot \|E_2\|_\infty \\
&\leq \|E_1\|_\infty \cdot \|\mathbf{fl}(\bar{B}^{(k+1)})\|_\infty + \sqrt{\alpha} \cdot \|E_2\|_\infty + O(\epsilon^2) \\
&\leq \eta_1 \alpha n \|\mathbf{fl}(\bar{A}^{(k+1)})\|_\infty \cdot \|\mathbf{fl}(\bar{B}^{(k+1)})\|_\infty + \eta_2 \alpha^{\frac{3}{2}} \cdot \|R\|_\infty \cdot \|\mathbf{fl}(\bar{B}^{(k+1)})\|_\infty + O(\epsilon^2) \\
&= \alpha \cdot \left(\eta_1 \cdot n \cdot \|\mathbf{fl}(\bar{A}^{(k+1)})\|_\infty + \eta_2 \cdot \sqrt{\alpha} \cdot \|R\|_\infty \right) \cdot \|\mathbf{fl}(\bar{B}^{(k+1)})\|_\infty + O(\epsilon^2) \\
&\leq \alpha \cdot \left(2\eta_1 \cdot n \cdot \sqrt{\alpha} + \eta_2 \cdot \alpha \cdot (\sqrt{n}+1) \right) \cdot \left(\sqrt{\alpha} \cdot n \cdot \mu \cdot \xi_{\max} + \sqrt{n} \cdot \|\hat{U}\|_\infty \cdot \xi_{\max} \right) + O(\epsilon^2) \\
&\leq 4\bar{\eta} \cdot (\alpha \cdot n)^{\frac{3}{2}} \cdot (n \cdot \mu + \|\hat{U}\|_\infty) \cdot \xi_{\max} + O(\epsilon^2) \,,
\end{aligned}$$

where $\bar{\eta} = \max\{\eta, \eta_1, \eta_2\}$, and we have used the fact that $\alpha \leq n$. Hence,

$$\|\hat{A}^{(k+1)} \cdot \hat{B}^{(k+1)} - \mathbf{fl}(\bar{A}^{(k+1)}) \cdot \mathbf{fl}(\bar{B}^{(k+1)})\|_\infty \leq 4\bar{\eta} \cdot (\alpha \cdot n)^{\frac{3}{2}} \cdot (n \cdot \mu + \|\hat{U}\|_\infty) \cdot \xi_{\max} + O(\epsilon^2).$$

In addition, a derivation similar to that for (5.16) gives

$$\|\mathbf{fl}(\bar{A}^{(k+1)}) \cdot \mathbf{fl}(\bar{B}^{(k+1)}) - \bar{A}^{(k+1)} \cdot \bar{B}^{(k+1)}\|_\infty \leq 4n\eta\mu(\nu + \psi\xi_{\max}) + O(\epsilon^2).$$

Combining these two relations we get

$$\begin{aligned}
\|\hat{G}^{(k+1)} - \bar{G}^{(k+1)}\|_\infty &= \|\hat{A}^{(k+1)} \cdot \hat{B}^{(k+1)} - \bar{A}^{(k+1)} \cdot \bar{B}^{(k+1)}\|_\infty \\
&\leq \|\hat{A}^{(k+1)} \cdot \hat{B}^{(k+1)} - \mathbf{fl}(\bar{A}^{(k+1)}) \cdot \mathbf{fl}(\bar{B}^{(k+1)})\|_\infty \\
&\quad + \|\mathbf{fl}(\bar{A}^{(k+1)}) \cdot \mathbf{fl}(\hat{B}^{(k+1)}) - \bar{A}^{(k+1)} \cdot \bar{B}^{(k+1)}\|_\infty \\
&\leq 4\bar{\eta} \cdot (\alpha \cdot n)^{\frac{3}{2}} \cdot (n \cdot \mu + \|\hat{U}\|_\infty) \cdot \xi_{\max} + 4\bar{\eta} n \mu (\nu + \psi \xi_{\max}) + O(\epsilon^2).
\end{aligned}$$

According to definition (5.4), this implies that for Algorithm 2

$$\tau \leq 4\bar{\eta} \cdot (\alpha \cdot n)^{\frac{3}{2}} \cdot (n \cdot \mu + \|\hat{U}\|_\infty) \cdot \xi_{\max} + 4\bar{\eta} n \mu (\nu + \psi \xi_{\max}) + O(\epsilon^2). \tag{5.23}$$

THEOREM 5.4. *For Algorithm* 2*, the backward error matrix* H *in* (5.1) *satisfies*

$$\|H\|_\infty \leq 8\sqrt{\alpha}\bar{\eta}(\alpha+2) \cdot n^2 \cdot \rho \cdot \left(\|\hat{U}\|_\infty + n \cdot \mu \right) + O(\epsilon^2),$$

where ρ *is defined in* (2.4).

Proof. We first derive upper bounds for ν and ψ, and then finish the proof by plugging relation (5.23) and these bounds into (5.13). Since $\hat{A}^{(k)}$ is numerically column orthogonal,

$$|\hat{a}_i^{(k)}|^T \cdot |\hat{b}_k^{(k)}| \leq \|\hat{a}_i^{(k)}\|_2 \cdot \|\hat{b}_k^{(k)}\|_2 \leq \|\hat{b}_k^{(k)}\|_2 + O(\epsilon).$$

We combine this and (5.21) to get

$$\frac{|\hat{a}_i^{(k)}|^T \cdot |\hat{b}_k^{(k)}|}{|\hat{\gamma}_k|} \leq \sqrt{n} \cdot \xi_{\max} + O(\epsilon).$$

By definition (5.5) this implies

$$\nu \leq \sqrt{n} \cdot \xi_{\max} + O(\epsilon). \tag{5.24}$$

On the other hand,

$$\begin{aligned}\left\||\hat{A}^{(k)}|\cdot|\hat{B}^{(k)}|\right\|_{\max} &\le \left\||\hat{A}^{(k)}|\cdot|\hat{B}^{(k)}|\right\|_2 \le \||\hat{A}^{(k)}|\|_2\cdot\||\hat{B}^{(k)}|\|_2\\ &\le \|\hat{A}^{(k)}\|_F\cdot\|\hat{B}^{(k)}\|_F = \sqrt{\alpha}\|\hat{B}^{(k)}\|_F + O(\epsilon),\end{aligned}$$

and

$$\left\|\hat{A}^{(k)}\cdot\hat{B}^{(k)}\right\|_{\max} \ge \frac{1}{n}\left\|\hat{A}^{(k)}\cdot\hat{B}^{(k)}\right\|_F = \frac{1}{n}\|\hat{B}^{(k)}\|_F + O(\epsilon).$$

Consequently,

$$\frac{\left\||\hat{A}^{(k)}|\cdot|\hat{B}^{(k)}|\right\|_{\max}}{\left\|\hat{A}^{(k)}\cdot\hat{B}^{(k)}\right\|_{\max}} \le \sqrt{\alpha}\cdot n + O(\epsilon).$$

By definition (5.5) this implies

$$\psi \le \sqrt{\alpha}\cdot n + O(\epsilon). \tag{5.25}$$

Plugging relations (5.23) through (5.25) into (5.13) we get

$$\begin{aligned}\|H\|_\infty &\le \frac{\alpha\eta(\nu+\psi\cdot\xi_{\max})}{\xi_{\min}}\cdot 2n\cdot\left(\|\hat{U}\|_\infty + n\cdot\mu\right) + \frac{n\tau}{\xi_{\min}} + O(\epsilon^2)\\ &\le \frac{2n\bar{\eta}(\alpha+2)(\nu+\psi\cdot\xi_{\max})}{\xi_{\min}}\cdot\left(\|\hat{U}\|_\infty + n\cdot\mu\right)\\ &\quad + \frac{4\bar{\eta}\cdot(\alpha\cdot n)^{\frac{3}{2}}}{\xi_{\min}}\cdot\left(\|\hat{U}\|_\infty + n\cdot\mu\right) + O(\epsilon^2)\\ &\le \frac{2n\bar{\eta}(\alpha+2)(\sqrt{n}\cdot\xi_{\max} + \sqrt{\alpha}\cdot n\cdot\xi_{\max})}{\xi_{\min}}\cdot\left(\|\hat{U}\|_\infty + n\cdot\mu\right)\\ &\quad + \frac{4\bar{\eta}\cdot(\alpha\cdot n)^{\frac{3}{2}}}{\xi_{\min}}\cdot\left(\|\hat{U}\|_\infty + n\cdot\mu\right) + O(\epsilon^2)\\ &\le 8\sqrt{\alpha}\bar{\eta}\cdot(\alpha+2)\cdot n^2\cdot\rho\cdot\left(\|\hat{U}\|_\infty + n\cdot\mu\right) + O(\epsilon^2). \qquad \square\end{aligned}$$

Remark 6. More detailed error analysis shows that the $O(\epsilon^2)$ term in Theorem 5.4 is bounded by $p_1(n)\epsilon$ times the first term, where $p_1(n)$ is a low-degree polynomial in n.

Remark 7. Throughout this analysis, we never used the fact that Algorithm 2 performs pivoting on the columns as well. Hence, Theorem 5.4 still holds if Algorithm 2 is modified to only perform partial pivoting.

Remark 8. An upper bound similar to that in Theorem 5.4 holds for Algorithm 3 as well, provided that

$$\||\hat{A}^{(k)}|\cdot|\hat{B}^{(k)}|\|_2 \approx \|\hat{A}^{(k)}\cdot\hat{B}^{(k)}\|_2$$

for all k, as is the case in our numerical experiments.

Remark 9. If C is transformed from a Toeplitz-plus-Hankel matrix via equation (3.3), then $\rho = O(n^2)$. In this case, the upper bound in Theorem 5.4 is a factor of $O(\alpha^{\frac{3}{2}}\, n^2)$ larger than the upper bound for the backward error in straightforward

GEPP and GECP, which is about $5\epsilon \cdot n^2 \cdot \|\hat{U}\|_\infty$ (see [15, p. 115]). Our numerical experiments indicate that for such matrices, Algorithm 2 is sometimes less accurate than straightforward GEPP, and the lost accuracy can be recovered by one step of iterative refinement [15, section 3.5]. See section 4 for more details.

Finally, we perform a brief error analysis for Algorithm 4. Let M be a Toeplitz-plus-Hankel-like matrix satisfying (3.2) with A column orthogonal. In finite arithmetic, Algorithm 4 factorizes M by performing the following computations.

- Compute $\hat{\Omega} = \mathbf{fl}(\mathcal{D}_{1,1})$ and $\hat{\Lambda} = \mathbf{fl}(\mathcal{D}_{1,-1})$.
- Compute $\hat{\mathcal{A}} = \mathbf{fl}(\mathcal{Q}_{1,1}^T \cdot A)$ and $\hat{\mathcal{B}} = \mathbf{fl}(B \cdot \mathcal{Q}_{1,-1})$.
- Compute the LU factorization for $\hat{C}$, where $\hat{C}$ is the Cauchy-like matrix that satisfies the displacement equation

$$\hat{\Omega} \cdot \hat{C} - \hat{C} \cdot \hat{\Lambda} = \hat{\mathcal{A}} \cdot \hat{\mathcal{B}}.$$

It is easy to show that

$$\|\mathcal{Q}_{1,1}^T \cdot M \cdot \mathcal{Q}_{1,-1} - \hat{C}\|_\infty \leq \eta_3 \cdot p_2(n) \cdot \|M\|_\infty,$$

where η_3 is a small multiple of ϵ and $p_2(n)$ is a low-degree polynomial in n. Thus the reduction from a Toeplitz-plus-Hankel-like matrix satisfying (3.2) to a Cauchy-like matrix satisfying (3.3) is numerically stable. In other words, Algorithm 4 is numerically stable if and only if the algorithm it uses in step 2 (Algorithm 1, 2, or 3) is stable.

6. Conclusions and extensions. We have presented a fast algorithm for solving Toeplitz or Toeplitz-plus-Hankel systems of linear equations and shown it to be numerically stable, provided that the element growth in the computed factorization is not large. We have presented practical modifications to this algorithm and discussed implementation techniques that further improve its efficiency. Our numerical experiments show that the resulting algorithm is both stable and efficient; and the cost for performing pivoting for Cauchy-like matrices can be kept a small fraction of the total cost.

The algorithms presented in this paper can be modified to solve *mosaic Toeplitz* or *block Toeplitz* systems of linear equations (see [7, 11]).

Our techniques to avoid internal element growth in the generators can be easily extended to the *generalized Schur algorithm* for factorizing more generally structured matrices (see [6, 22, 23]), and so is the technique to store the rows of the U matrix columnwise.

Recently, Chandrasekaran and Sayed [5] proposed a new fast algorithm for factorizing the Toeplitz matrix based on the QR factorization of a larger structured matrix and show that it is numerically stable. This algorithm appears to perform more flops than Algorithm 4 but does not have the potential problem of having large element growth in the computed factorization.

We end this paper by asking two open questions.

1. The upper bounds for element growth on GEPP and the variation of GECP in Algorithm 2 on a Cauchy-like matrix are 2^{n-1} and $\rho^{2+\sum_{k=1}^{n-1} 1/k} \cdot \mathcal{W}(n)$, respectively (see section 2). Do sharper bounds exist for Cauchy-like matrices with low displacement rank?
2. There are superfast algorithms for solving Toeplitz or Toeplitz-plus-Hankel systems of linear equations in $O(n \log_2^2 n)$ flops, but they are unstable in

general (see [3]). Are there numerically stable superfast algorithms for such problems?

Acknowledgments. The author is grateful to Profs. S. Chandrasekaran, J. Demmel, and Dr. V. Olshevsky for helpful discussions, and Profs. S. Chandrasekaran, I. Gohberg, G. Heinig, T. Kailath, A. H. Sayed, and Dr. V. Olshevsky for preprints of their papers.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, 2nd ed., SIAM, Philadelphia, PA, 1994.

[2] E. Bozzo and C. Di Fiore, *On the use of certain matrix algebras associated with discrete trigonometric transforms in matrix displacement decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 312–326.

[3] J. R. Bunch, *Stability of methods for solving Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 349–364.

[4] T. F. Chan and P. C. Hansen, *A lookahead Levinson algorithm for general Toeplitz systems*, IEEE Proc. Signal Processing, 40 (1992), pp. 1079–1090.

[5] S. Chandrasekaran and A. H. Sayed, *A fast stable solver for nonsymmetric Toeplitz and quasi-Toeplitz systems of linear equations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 107–139.

[6] S. Chandrasekaran and A. H. Sayed, *Stabilizing the fast generalized Schur algorithm*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 950–983.

[7] J. Chun and T. Kailath, *Generalized displacement structure for block-Toeplitz, Toeplitz-block, and Toeplitz-derived matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 114–128.

[8] G. Cybenko, *The numerical stability of Levinson-Durbin algorithm for Toeplitz systems of equations*, SIAM J. Sci. Statist. Comput., 1 (1980), pp. 303–319.

[9] W. J. Demmel, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.

[10] M. Fiedler, *Hankel and Loewner matrices*, Linear Algebra Appl., 58 (1984), pp. 75–95.

[11] I. Gohberg, T. Kailath, and V. Olshevsky, *Fast Gaussian elimination with partial pivoting for matrices with displacement structure*, Math. Comp., 64 (1995), pp. 1557–1576.

[12] I. Gohberg and V. Olshevsky, *Complexity of multiplication with vectors for structured matrices*, Linear Algebra Appl., 202 (1994), pp. 163–192.

[13] I. Gohberg and V. Olshevsky, *Fast algorithm for matrix Nehari problem*, in Systems and Networks: Mathematical Theory and Applications, Proc. of the International Symposium MTNS-93, U. Helmke, R. Mennicken, and J. Sauers, eds., Akademie Verlag, Berlin, 2 (1994), pp. 687–690.

[14] I. Gohberg and V. Olshevsky, *Fast state space algorithms for matrix Nehari and Nehari-Takagi interpolation problems*, Integral Equations Oper. Theory, 20 (1994), pp. 44–83.

[15] G. Golub and C. van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[16] M. H. Gutknecht and Hochbruck M., *Look-Ahead Levinson and Schur Algorithms for Non-Hermitian Toeplitz Systems*, IPS Research Report 93-11, IPS Supercomputing, ETH-Zürich, Switzerland, 1993.

[17] G. Heinig, *Inversion of generalized Cauchy matrices and other classes of structured matrices*, in Linear Algebra in Signal Processing, IMA Vol. Math. Appl., 69 (1994), pp. 95–114.

[18] G. Heinig, P. Jankowski, and K. Rost, *Fast inversion of Toeplitz-plus-Hankel matrices*, Numer. Math, 52 (1988), pp. 665–682.

[19] G. Heinig and K. Rost, *Algebraic methods for Toeplitz-like matrices and operators*, Oper. Theory, 13 (1984), pp. 109–127.

[20] A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Dover, New York, 1964.

[21] T. Kailath, S. Kung, and M. Morf, *Displacement ranks of matrices and linear equations*, J. Math. Anal. Appl., 68 (1979), pp. 395–407.

[22] T. Kailath and A. H. Sayed, *Fast algorithms for generalized displacement structures*, in Recent Advances in Mathematical Theory of Systems, Control, Networks, and Signal Processing, Vol. II, H. Kimura and S. Kodama, eds., Mita Press, Japan, 1992, pp. 27–32.

[23] T. Kailath and A. H. Sayed, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.
[24] V. Pan, *On computations with dense structured matrices*, Math. Comp., 55 (1990), pp. 179–190.
[25] D. R. Sweet, *The use of pivoting to improve the numerical performance of Toeplitz matrix algorithms*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 468–493.
[26] D. R. Sweet and R. P. Brent, *Error analysis of a fast partial pivoting method for structured matrices*, in Adv. Signal Proc. Algorithms, Proc. of SPIE, T. Luk, ed., 2363 (1995), pp. 266–280.
[27] J. M. Varah, *The Prolate matrix*, Linear Algebra Appl., 187 (1993), pp. 269–278.
[28] J. H. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 10 (1961), pp. 281–330.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 307–324, April 1998

002

H-SELFADJOINT AND H-UNITARY MATRIX PENCILS*

ILYA KRUPNIK† AND PETER LANCASTER†

Abstract. A square matrix pencil $\lambda A - B$ is said to be H-selfadjoint (H-unitary) if it satisfies $A^*HB = B^*HA$ ($A^*HA = B^*HB$) for some invertible Hermitian H. Attention is focused on regular pencils (i.e., $\det(\lambda A - B) \not\equiv 0$) for which A and B are both singular. Canonical forms for the relation $(A, B, H) \sim (Y^{-1}AX, Y^{-1}BX, Y^*HY)$ are obtained in both the complex and real cases. Also, a characterization is given for those real matrices A which are H-unitary for some H, i.e., $A^THA = H$ for some invertible, real symmetric H.

Key words. symmetric pencils, canonical forms

AMS subject classifications. 15A21, 15A57

PII. S0895479895286025

1. Introduction.

In this paper we undertake a study of canonical forms for regular matrix pencils $\lambda A - B$ (i.e., $\det(\lambda A - B) \not\equiv 0$) which are either H-selfadjoint or H-unitary. These terms are defined as follows:

(a) A regular $n \times n$ pencil $\lambda A - B$ is said to be *H-selfadjoint* if H is an $n \times n$ nonsingular Hermitian matrix and

$$(A^*HB)^* = A^*HB. \tag{1.1}$$

(b) A regular $n \times n$ pencil $\lambda L - M$ is said to be *H-unitary* if H is an $n \times n$ nonsingular Hermitian matrix and

$$L^*HL = M^*HM. \tag{1.2}$$

Pencils of these two varieties arise in the study of minimal realizations of rational matrix functions in the form $W(\lambda) = D(\lambda A - B)^{-1}C$ when $W(\lambda)$ is Hermitian on the real line (case (1.1)) and when $W(\lambda)$ is Hermitian on the unit circle (case (1.2)), (see [8]). Pencils satisfying (1.2) also arise in the study of so-called "discrete algebraic Riccati equations" with a special choice of H, and were studied by Wimmer [13] in 1991. See Chapter 15 of [9] for a general treatment. The relationship between pencils of the two kinds has, however, been recognized in this context for some time. In the work of Gardiner and Laub [4] in 1986, for example, good use is made of the Cayley transform technique to transform from one to the other. The "Hamiltonian" and "symplectic" pencils discussed in those papers are included in our analysis. A start is made on a more wide-ranging discussion of H-selfadjoint and H-unitary pencils by Lancaster and Rodman [9], and it is our purpose to continue with the analysis initiated there.

Let us briefly review some terminology used in the sequel. A number $\lambda_0 \in \mathbb{C}$ is an *eigenvalue* of pencil $\lambda A - B$ if $\lambda_0 A - B$ is singular. The set of all eigenvalues of a pencil is known as its *spectrum* and is written $\sigma(A, B)$. As all pencils in this paper are regular, the spectrum is a finite subset of $\mathbb{C}$ and contains the point at infinity when A is singular.

*Received by the editors May 12, 1995; accepted for publication (in revised form) by P. Van Dooren January 14, 1997. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/simax/19-2/28602.html

†Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta T2N 1N4, Canada (lancaste@acs.ucalgary.ca).

A subspace $\mathcal{S}$ of $\mathbb{C}^n$ is said to be *deflating* for the regular pencil $\lambda A - B$ if there is a subspace $\mathcal{T}$ such that $\dim \mathcal{S} = \dim \mathcal{T}$, $A\mathcal{S} \subseteq \mathcal{T}$, and $B\mathcal{S} \subseteq \mathcal{T}$. (In particular, when $A = I$, $\mathcal{S}$ is B-invariant.) A deflating subspace is said to be *spectral* if

$$\mathcal{S} = \operatorname{Im} \int_\Gamma (\lambda A - B)^{-1} A \, d\lambda$$

for some simple closed curve Γ which does not intersect $\sigma(\lambda A - B)$ (see the Appendix for further discussion).

Now let us consider some special cases of our problem. First, and most importantly, consider the case $A = I$. Then $\lambda A - B$ is H-selfadjoint (H-unitary) means that $B^*H = HB$ (or $H = B^*HB$) and B is said to be an *H-selfadjoint* (or *H-unitary*) matrix. Matrices of these kinds have been systematically studied in [7] and we rely heavily on that work. Thus, when B is H-selfadjoint we seek to reduce the pair (H, B) *simultaneously* by a congruence and a similarity.

$$X^*HX = P_{\varepsilon,J}, \qquad X^{-1}BX = J$$

or, which is equivalent, we reduce the Hermitian matrices H and HB by a simultaneous congruence,

$$X^*HX = P_{\varepsilon,J}, \qquad X^*(HB)X = P_{\varepsilon,J}J.$$

The precise description of the canonical matrices $P_{\varepsilon,J}$ and (Jordan canonical form) J can be found in Theorem I.4.1 of [7] (see also Chapter S5 of [6]). We refer to $(P_{\varepsilon,J}, J)$ as a (complex) *selfadjoint canonical pair* for (H, B). Let us just recall that, as well as the usual similarity invariants implicit in J, $P_{\varepsilon,J}$ encodes information on the *sign characteristic* ε of B with respect to H (a set of $+1$'s and -1's associated with the partial multiplicities of real eigenvalues of B).

Similarly, if B if H-unitary, there is a transforming matrix X such that

$$X^*HX = Q_{\varepsilon,J}, \qquad X^{-1}BX = J,$$

and $(Q_{\varepsilon,J}, J)$ is a (complex) *unitary canonical pair* for (H, B) and they are described in Theorem I.4.3 of [7] (see also [9]).

When B and H are real matrices and real canonical forms are required, we can again refer to [7]. See Theorem I.5.3 for the description of *real selfadjoint canonical pairs* $(P_{\varepsilon,J}, J)$. To our knowledge, a complete description of *real unitary canonical pairs* has not been written down to date and so we avoid the use of this phrase. (The complex case (Theorem I.4.3 of [7]) is difficult enough and, in principle, the real unitary case can always be obtained using the techniques of [7]. The difficulty lies in the fact that the "simplest" member of an equivalence class is still complicated, and the specification of a "simplest" member is not obvious.)

If the pencil $\lambda A - B$ is an H-selfadjoint pencil and either A or B is nonsingular, there is an underlying H-selfadjoint matrix. For example, if A^{-1} exists and (1.1) holds, then $H(BA^{-1}) = (BA^{-1})^*H$. For this reason, our attention is focused on regular pencils for which both A and B are singular, and similarly for the unitary case.

Another important and well-understood case is that of a selfadjoint pencil $\lambda A - B$, i.e., when $A^* = A$ and $B^* = B$. See Theorems I.3.18 and I.5.4 of [7] for the complex and real cases, respectively. They were also studied by Elsner and Lancaster [3], and

were utilized in the work of Clements and Glover [1], for example. In this case we have the following.

PROPOSITION 1.1. *If $A^* = A$, $B^* = B$, and $\lambda A - B$ is regular, then there is a nonsingular Hermitian matrix H such that* (1.1) *holds.*

Proof. Choose a $\lambda_0 \in \mathbb{R}$ such that $\lambda_0 A - B$ is nonsingular and define $H = (\lambda_0 A - B)^{-1}$. Then $H^* = H$ and

$$A = AHH^{-1} = \lambda_0 AHA - AHB.$$

Since A and $\lambda_0 AHA$ are Hermitian, so are AHB and A^*HB. □

The analysis of this paper is based on strict-equivalence transformations of pencils: $\lambda A - B \to Y^{-1}(\lambda A - B)X$, where X is also nonsingular. The selfadjoint or unitary property of the pencil is preserved in the following sense.

PROPOSITION 1.2. *If $\lambda A - B$ is H-selfadjoint (or H-unitary) and*

$$Y^{-1}(\lambda A - B)X = \lambda \hat{A} - \hat{B}, \tag{1.3}$$

then $\lambda \hat{A} - \hat{B}$ is $\hat{H}$-selfadjoint (or $\hat{H}$-unitary, respectively), where $\hat{H} = Y^*HY$.

The proof is a simple verification.

Suppose we are dealing with matrices of a fixed size n and define

$$\mathcal{U} = \{(A, B, H) : \det(\lambda A - B) \not\equiv 0,\ H^* = H,\ \det H = 0\}.$$

We say that $(A_1 B_1, H_1)$, (A_2, B_2, H_2) from $\mathcal{U}$ are *unitarily equivalent* if there exist nonsingular X and Y such that

$$A_1 = Y^{-1}A_2X, \quad B_1 = Y^{-1}B_2X, \quad H_1 = Y^*H_2Y,$$

(cf. Section 3.1 of [7]). It is easily verified that unitary equivalence is an equivalence relation on $\mathcal{U}$. Thus, we seek a canonical form characterizing the corresponding equivalence classes.

Section 2 concerns some preliminaries on H-selfadjoint pencils and a careful discussion of Cayley transforms appears in section 3. Canonical forms over $\mathbb{C}$ are obtained in section 4. Although the canonical form for H-unitary pencils (Theorem 4.2) could be obtained directly, we prefer to use the Cayley transform and derive this from the H-selfadjoint case of Theorem 4.1.

Sections 5 and 6 concern the corresponding forms over the real numbers and show that apart from the obvious changes in real Jordan structure, there are no other essential differences in the canonical forms for the complex and real cases.

Section 7 concerns a problem of a related but different kind. A characterization is given of those real matrices A which are H-unitary for some real, symmetric, and nonsingular H, i.e., for which $A^THA = H$ for some such H. This is in contrast to the relatively well-known H-selfadjoint case. For example, Corollary I.5.2 of [7] asserts that *every* real square matrix A is H-selfadjoint for some invertible, real symmetric matrix H. It turns out (see [8]) that this problem applies to the real realization of matrix functions $W(\lambda)$ which are Hermitian on the unit circle.

2. Preliminaries on H-selfadjoint pencils. Let us begin with another characterization of regular H-selfadjoint matrix pencils.

THEOREM 2.1. *Let $\lambda A - B$ be a regular pencil. Then $\lambda A - B$ is H-selfadjoint if and only if TA and TB are Hermitian, where $T = (\mu A^* - B^*)H$ and μ is any real number for which $\mu A - B$ is nonsingular.*

Proof. Let $\lambda A - B$ be H-selfadjoint and use the definition $A^*HB = B^*HA$ to verify that, for any $\lambda, \mu \in \mathbb{C}$

$$(\lambda A^* - B^*)H(\mu A - B) = (\mu A^* - B^*)H(\lambda A - B).$$

Now choose a fixed $\mu \in \mathbb{R}$ for which $\mu A - B$ is nonsingular and define $T = (\mu A^* - B^*)H$. Then the last displayed equation gives

$$(\lambda A^* - B^*)T^* = T(\lambda A - B) \tag{2.1}$$

for all $\lambda \in \mathbb{C}$ and hence TA and TB are Hermitian.

Conversely, we have TA and TB are Hermitian and $H = T^*(\mu A - B)^{-1}$. Then we may write

$$H = T^*(\mu TA - TB)^{-1}T$$

which shows that H is nonsingular and Hermitian. Finally, because TB is Hermitian so is

$$(\mu A^* - B^*)HB = \mu A^*HB - B^*HB,$$

and it follows that A^*HB is Hermitian, as required. □

Observe now that (2.1) yields a strict equivalence and implies that $\lambda A - B$ and $\lambda A^* - B^*$ have the same spectra and partial multiplicities. Consequently, $\sigma(A, B)$ is symmetric with respect to the real line. Also, as observed in Proposition 1.1, when $A^* = A$ and $B^* = B$ we may take $H = (\mu A - B)^{-1}$ and $T = I$.

Combining Proposition 1.2 and Theorem 2.1 we have the following.

COROLLARY 2.2. *If $\lambda A - B$ is H-selfadjoint, then $\lambda A^* - B^*$ is $\hat{H}$-selfadjoint, where $\hat{H} = WHW^*$ and*

$$W = (T^{-1})^* = (H(\mu A - B))^{-1}. \tag{2.2}$$

From Theorem 2.1 it can also be deduced that, for any H-selfajoint pencil $\lambda A - B$, the spectrum and systems of *right* eigenvectors and generalized eigenvectors are just those of a regular Hermitian pencil $\lambda TA - TB$.

Note that although $\hat{H}$ depends on the choice of μ in (2.2), the signature of $\hat{H}$ always agrees with that of H. Also, we may write

$$\hat{H}^{-1} = (\mu A^* - B^*)H(\mu A - B). \tag{2.3}$$

As an alternative to expressing $\hat{H}$ in terms of μ it is possible to determine $\hat{H}$ in terms of a deflation of $\lambda A - B$ (see the Appendix), as described in the following.

PROPOSITION 2.3. *If $\lambda A - B$ is H-selfadjoint and X, Y have the properties that Y is nonsingular and*

$$Y^{-1}(\lambda A - B)X = \begin{bmatrix} \lambda I - T_1 & 0 \\ 0 & \lambda T_2 - I \end{bmatrix}, \tag{2.4}$$

where $Y = [Y_1 \;\; Y_2]$ and Y_1, T_1 have the same number of columns, then, in Corollary 2.2, we may take

$$W = \begin{bmatrix} Y_1^*HA \\ Y_2^*HB \end{bmatrix}^{-1} Y^*. \tag{2.5}$$

Proof. This follows from the proof of Theorem 2.8.1 of [9]. ∎

THEOREM 2.4. *Let $\lambda A - B$ be H-selfadjoint, let (2.4) hold, and assume that $\sigma(\lambda I - T_1^*) \cap \sigma(\lambda T_2 - I) = \emptyset$. Then $\lambda I - T_1$ and $\lambda T_2 - I$ are H_1-selfadjoint and H_2-selfadjoint, respectively, where $H_1 = Y_1^* H Y_1$, $H_2 = Y_2^* H Y_2$. Furthermore,*

$$Y^* H Y = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix}. \tag{2.6}$$

Proof. By Proposition 1.2 the pencil on the right of (2.4) is Y^*HY-selfadjoint. This means that if we write $H_3 = Y_1^* H Y_2$, then

$$\begin{bmatrix} I & 0 \\ 0 & T_2^* \end{bmatrix} \begin{bmatrix} H_1 & H_3 \\ H_3^* & H_2 \end{bmatrix} \begin{bmatrix} T_1 & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} T_1^* & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} H_1 & H_3 \\ H_3^* & H_2 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & T_2 \end{bmatrix}.$$

This is equivalent to the three relations

$$T_1^* H_1 = H_1 T_1, \qquad T_2^* H_2 = H_2 T_2, \tag{2.7}$$

and $H_3 - T_1^* H_3 T_2 = 0$. Since $\sigma(\lambda I - T_1^*) \cap \sigma(\lambda T_2 - I) = \emptyset$ it follows from Theorem A.4 that $H_3 = 0$ and (2.6) holds. Since H is nonsingular, so are H_1 and H_2 and it follows from (2.7) that T_j is H_j-selfadjoint for $j = 1$ and 2. ∎

Let us write $X = [X_1 \;\; X_2]$, as in Proposition A.3. Then $Y_1 = AX_1$ and $Y_2 = BX_2$. Thus, the inner products in which $\lambda I - T_1$ and $\lambda T_2 - I$ are selfadjoint are defined by

$$X_1^*(A^* H A) X_1, \qquad X_2^*(B^* H B) X_2, \tag{2.8}$$

respectively, and (see Theorem A.2 and (A.1)) $\operatorname{Im} X_1$ and $\operatorname{Im} X_2$ are the image and kernel of the projection P, respectively.

Notice also that if the reduced pencil of (2.4) is obtained as described in Theorem A.2, then $\sigma(\lambda I - T_1) \cap \sigma(\lambda T_2 - I) = \emptyset$ as well. Clearly, $\det T_1 = 0$ if and only if B is singular and $\det T_2 = 0$ if and only if A is singular. It will be convenient for us to suppose that (2.4) is obtained using Theorem A.2 and letting Γ contain *all* the finite spectrum of $\lambda A - B$. Then $\sigma(\lambda T_2 - I) = \{\infty\}$ (assuming that A is singular).

3. Cayley transforms. As in classical matrix theory, fractional linear transformations of the form $\lambda = (\alpha\bar{w} - w\mu)/(\alpha - \mu)$, $\bar{w} \neq w$, $|\alpha| = 1$, can be used to map H-selfadjoint pencils onto H-unitary pencils as follows.

PROPOSITION 3.1.

(a) *If $\lambda A - B$ is an H-selfadjoint pencil and $w, \alpha \in \mathbb{C}$ with $w \neq \overline{w}$, $|\alpha| = 1$, and if*

$$L = wA - B, \qquad M = \alpha(\overline{w}A - B), \tag{3.1}$$

then $\lambda L - M$ is an H-unitary pencil.

(b) *If $\lambda L - M$ is an H-unitary pencil, $w \neq \overline{w}$, $|\alpha| = 1$, and if*

$$A = \alpha L - M, \qquad B = \overline{w}\alpha L - wM, \tag{3.2}$$

then $\lambda A - B$ is an H-selfadjoint pencil.

The proof is a straightforward verification. An analogue of Theorem 2.1 follows.

THEOREM 3.2. *If $\lambda L - M$ is a regular H-unitary pencil, then there is a nonsingular W such that*

$$(W^*)^{-1}(\lambda L - M)W = \lambda M^* - L^*. \tag{3.3}$$

Proof. Define A and B as in (3.2). Then, by Proposition 3.1, $\lambda A - B$ is H-selfadjoint and, by Theorem 2.1, there is a W (take $W = (T^*)^{-1}$) such that

$$(W^*)^{-1}AW = A^*, \qquad (W^*)^{-1}BW = B^*.$$

Substitute for A and B from (3.2) and, using $w \neq \overline{w}$, it follows that

$$(W^*)^{-1}MW = -\alpha L^*, \qquad (W^*)^{-1}LW = -\alpha^{-1}M^*.$$

Now choose $\alpha = -1$ to obtain the result. ∎

As (3.3) is a strict equivalence, it follows readily that the spectrum of a regular H-unitary pencil is symmetric with respect to the unit circle. Notice in particular that in contrast with H-unitary matrices, H-unitary pencils may have a zero eigenvalue. Combining the symmetry of the spectrum with Proposition 3.1, the following proposition is readily proved and provides a convenient tool for the study of H-unitary pencils.

PROPOSITION 3.3.

(a) *Let $\lambda A - B$ be a regular H-selfadjoint pencil and $w \in \sigma(\lambda A - B)$, with $w \neq \overline{w}$. Then the pencil $\lambda L - M$ defined by (3.1) is H-unitary and has eigenvalues at zero and infinity. The partial multiplicities of these two eigenvalues agree and are equal to those of w (and of $\overline{w}$) as an eigenvalue of $\lambda A - B$.*

(b) *If $\lambda L - M$ is a regular H-unitary pencil, then $O \in \sigma(\lambda L - M)$ if and only if $\infty \in \sigma(\lambda L - M)$ and, in this case, these two eigenvalues have the same partial multiplicities. Also, if a pencil $\lambda A - B$ is defined by (3.2) with $w \neq \overline{w}$, $|\alpha| = 1$, then $\lambda A - B$ is a regular H-selfadjoint pencil and $w, \overline{w} \in \sigma(\lambda A - B)$ with partial multiplicities which agree with those of zero (or of ∞) as an eigenvalue of $\lambda L - M$.*

4. Canonical forms over $\mathbb{C}$. Let us first consider the canonical reduction of A, B, and H when $\lambda A - B$ is an H-selfadjoint pencil.

THEOREM 4.1. *If $\lambda A - B$ is a regular H-selfadjoint pencil, then there exist nonsingular matrices X and Y such that*

$$Y^{-1}(\lambda A - B)X = \begin{bmatrix} \lambda I - J & 0 \\ 0 & \lambda K - I \end{bmatrix}, \; Y^*HY = \begin{bmatrix} P_{\varepsilon_1, J} & 0 \\ 0 & P_{\varepsilon_2, K} \end{bmatrix}, \tag{4.1}$$

where $(P_{\varepsilon_1, J}, J)$ and $(P_{\varepsilon_2, K}, K)$ are selfadjoint canonical pairs and K is nilpotent.

Proof. By Theorem 2.4 there are nonsingular matrices X_0 and Y_0 such that

$$Y_0^{-1}(\lambda A - B)X_0 = \begin{bmatrix} \lambda I - T_1 & 0 \\ 0 & \lambda T_2 - I \end{bmatrix}, \qquad Y_0^* H Y_0 = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix},$$

where T_2 is nilpotent, H_1, H_2 are defined as in (2.8), and T_j is H_j-selfadjoint for $j = 1$ and 2.

Now form the canonical reductions

$$V_1^{-1}T_1V_1 = J, \quad V_1^*H_1V_1 = P_{\varepsilon_1,J},$$
$$V_2^{-1}T_2V_2 = K, \quad V_2^*H_2V_2 = P_{\varepsilon_2,K},$$

and the theorem is obtained by taking

$$X = X_0 \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}, \qquad Y = Y_0 \begin{bmatrix} V_1 & 0 \\ 0 & V_2 \end{bmatrix}. \qquad \square$$

It is natural now to refer to $\varepsilon := \{\varepsilon_{1,J}, \varepsilon_{2,K}\}$ as the sign characteristic of $\lambda A - B$ with respect to H, or of the triple $(A, B, H) \in \mathcal{U}$. The two matrices on the right of (4.1) constitute the canonical form and are determined by J, K, and ε. They are "canonical" in the sense that they characterize the equivalence class of $\mathcal{U}$ to which (A, B, H) belong. The canonical form is unique up to ordering of the Jordan blocks in J and K and the ordering of signs in ε corresponding to blocks of the same size with the same eigenvalue. The formal proof of these statements is an easy extension of the arguments in Section 3.5 of [7]. Similar remarks apply to Theorem 5.3 below.

THEOREM 4.2. *Let $\lambda L - M$ be a regular H-unitary pencil with a zero eigenvalue of algebraic multiplicity m; then there exist nonsingular matrices X and Y such that*

$$Y^{-1}(\lambda L - M)X = \begin{bmatrix} \lambda I - S & 0 & 0 \\ 0 & \lambda I - N & 0 \\ 0 & 0 & \lambda N^T - I \end{bmatrix}, \quad Y^*HY = \begin{bmatrix} G & 0 & 0 \\ 0 & 0 & -iI_m \\ 0 & iI_m & 0 \end{bmatrix}, \tag{4.2}$$

where (G, S) is a unitary canonical pair and N is a nilpotent Jordan matrix.

Proof. Without loss of generality it is assumed throughout that $m > 0$. Choose $\alpha, w \in \mathbb{C}$ with $\overline{w} \neq w$, $|\alpha| = 1$ and, for convenience, $\alpha(w - \overline{w}) = 1$. Let A, B be defined as in (3.2). Then, by Proposition 3.1, $\lambda A - B$ is an H-selfadjoint pencil. Furthermore, by Proposition 3.3, w, $\overline{w}$ are eigenvalues of $\lambda A - B$, each with algebraic multiplicity m.

Now use Theorem 2.4 (followed by a canonical reduction) to obtain nonsingular matrices X_0, Y_0 such that

$$Y_0^{-1}(\lambda A - B)X_0 = \begin{bmatrix} \lambda\hat{A} - \hat{B} & 0 \\ 0 & \lambda I_{2m} - \hat{J} \end{bmatrix}, \qquad Y_0^*HY_0 = \begin{bmatrix} G & 0 \\ 0 & \hat{P} \end{bmatrix},$$

where

$$\hat{J} = \begin{bmatrix} wI_m + N & 0 \\ 0 & \overline{w}I_m + N \end{bmatrix}, \qquad \hat{P} = \begin{bmatrix} 0 & P \\ P & 0 \end{bmatrix}.$$

Here, N is the $m \times m$ nilpotent Jordan matrix defined by the zero eigenvalue of $\lambda L - M$, $(\hat{P}, \hat{J})$ is a selfadjoint canonical pair, and $\lambda\hat{A} - \hat{B}$ is G-selfadjoint.

Now transform back to a unitary pencil $\lambda L_1 - M_1$ using (3.1) and, because $\alpha(w - \overline{w}) = 1$, we obtain $L_1 = Y_0^{-1}LX_0$, $M_1 = Y_0^{-1}MX_0$, and $\lambda L_1 - M_1$ is $\left[\begin{smallmatrix} \hat{H} & 0 \\ 0 & \hat{P} \end{smallmatrix}\right]$-unitary. Thus,

$$L_1 = \begin{bmatrix} w\hat{A} - \hat{B} & 0 \\ 0 & wI_{2m} - \hat{J} \end{bmatrix}, \qquad M_1 = \alpha \begin{bmatrix} \overline{w}\hat{A} - \hat{B} & 0 \\ 0 & \overline{w}I_{2m} - \hat{J} \end{bmatrix}.$$

Observe that if we define $J_0 = (\overline{w} - w)I_m + N$, then

$$wI_{2m} - \hat{J} = \begin{bmatrix} -N & 0 \\ 0 & -J_0 \end{bmatrix}, \qquad \overline{w}I_{2m} - \hat{J} = \begin{bmatrix} -\overline{J}_0 & 0 \\ 0 & -N \end{bmatrix}.$$

Multiply $\lambda L_1 - M_1$ on the right by $\operatorname{diag}[(w\hat{A} - \hat{B})^{-1}, -\overline{J}_0^{-1}, -J_0^{-1}]$ and we obtain an X_1 such that

$$Y_0^{-1}(\lambda L - M)X_1 = \begin{bmatrix} \lambda I - S & 0 & 0 \\ 0 & \lambda N\overline{J}_0^{-1} - I_m & 0 \\ 0 & 0 & \lambda I_m - NJ_0^{-1} \end{bmatrix} \tag{4.3}$$

and this pencil is $\left[\begin{smallmatrix} G & 0 \\ 0 & \hat{P} \end{smallmatrix}\right]$-unitary.

The block-Toeplitz structure of J_0^{-1} ensures that NJ_0^{-1} is similar to N. Let $N = T^{-1}(NJ_0^{-1})T$ and

$$V = \begin{bmatrix} 0 & P(T^*)^{-1} \\ iT & 0 \end{bmatrix}.$$

From the unitary property of (4.3) we deduce that $P(N\overline{J}_0^{-1}) = (NJ_0^{-1})^*P$ and hence that

$$T^*P(N\overline{J}_0^{-1})P(T^*)^{-1} = T^*(NJ_0^{-1})^*(T^*)^{-1} = N^T.$$

Using this fact it is easily verified that

$$V^{-1}\begin{bmatrix} \lambda N\overline{J}_0^{-1} - I & 0 \\ 0 & \lambda I - NJ_0^{-1} \end{bmatrix} V = \begin{bmatrix} \lambda I - N & 0 \\ 0 & \lambda N^T - I \end{bmatrix}.$$

Also, we have

$$V^*\hat{P}V = V^*\begin{bmatrix} 0 & P \\ P & 0 \end{bmatrix} V = \begin{bmatrix} 0 & -iI \\ iI & 0 \end{bmatrix}.$$

Now (4.3) is transformed to the desired form of (4.2). Clearly, it may also be assumed that G and S are in canonical form, and the theorem is proved. □

Remark. The canonical form presented in (4.2) is used mainly for historical reasons. It is most easily compared with Theorem 2.1 of Wimmer [13] (in which H has the form $(-i)\left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]$) and with the exposition of Lancaster and Rodman [9, Theorem 2.9.5 and sections 8.6, 15.1], where the form of H is inherited from formulation of properties of algebraic Riccati equations.

In other problem areas it may seem more natural to replace the second of equations (4.2) by

$$Y^*HY = \begin{bmatrix} G & 0 & 0 \\ 0 & 0 & I \\ 0 & I & 0 \end{bmatrix}. \tag{4.4}$$

This is easily achieved by postmultiplying X and Y by $\operatorname{diag}[I, -iI_m, I_m]$ and has the same form as the real case discussed below as Theorem 6.1.

Example. The pencil $\lambda\left[\begin{smallmatrix}1 & 0\\ 0 & 0\end{smallmatrix}\right] + \left[\begin{smallmatrix}0 & 0\\ 0 & 1\end{smallmatrix}\right]$ is selfadjoint with respect to $H = I$. In Theorem 4.1 we must have $P_J = P_K = 1$ and we may take $X = Y = I$. The pencil is also H-unitary with $H = \left[\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right]$. Then the first of equations (4.2) and equation (4.4) hold with $X = Y = I$ (and G, S do not appear).

5. Canonical forms for real H-selfadjoint pencils. Let us first summarize a statement of the real canonical form for real matrices under similarity (see [7], for example, for more details).

PROPOSITION 5.1. *If A is a real square matrix, then there exists a nonsingular real matrix Y such that*

$$Y^{-1}AY = \operatorname{diag}[K_r, K_c, K_0],$$

where K_r, K_c, K_0 are canonical real Jordan matrices with $\sigma(K_r) \subset \mathbb{R}\backslash\{0\}$, $\sigma(K_c) \subset \mathbb{C}\backslash\mathbb{R}$, and K_0 is nilpotent.

We use this to confirm that there is a real form of the Kronecker reduction for linear pencils.

PROPOSITION 5.2. *If $\lambda A - B$ is a real regular pencil, then there exist real nonsingular matrices X and Y such that*

$$Y^{-1}(\lambda A - B)X = \operatorname{diag}[\lambda I - J_r, \lambda I - J_c, \lambda J_0 - I],$$

where J_r, J_c, J_0 are real canonical Jordan matrices with the following properties: $\sigma(J_r) \subset \mathbb{R}$, $\sigma(J_c) \subset \mathbb{C}\backslash\mathbb{R}$, $\sigma(J_0) = \{0\}$.

Proof. Let $\lambda_0 \in \mathbb{R}$ with $\det(\lambda_0 A - B) \neq 0$, and define

$$C = -A(\lambda_0 A - B)^{-1}.$$

Apply Proposition 5.1 to C to obtain a real canonical form:

$$Y_0^{-1}CY_0 := K = \operatorname{diag}[K_r, K_c, K_0].$$

Note also that

$$I + \lambda_0 K = -Y_0^{-1}B(\lambda A - B)^{-1}Y_0. \tag{5.1}$$

Now set $X_0 = -(\lambda_0 A - B)^{-1}Y_0$ and we have the real strict equivalence

$$Y_0^{-1}(\lambda A - B)X_0 = \operatorname{diag}[\lambda K_r - (I + \lambda_0 K_r), \quad \lambda K_c - (I + \lambda_0 K_c), \quad \lambda K_0 - (I + \lambda_0 K_0)].$$

Furthermore, the right-hand matrix is strictly equivalent over $\mathbb{R}$ to

$$\operatorname{diag}[\lambda I - (I + \lambda_0 K_r)K_r^{-1}, \quad \lambda I - (I + \lambda_0 K_0)^{-1}, \quad \lambda K_0(I + \lambda_0 K_0)^{-1} - I]$$

and this, in turn, is strictly equivalent (in fact, similar) over $\mathbb{R}$ to

$$\operatorname{diag}[\lambda I - J_r, \lambda I - J_c, \lambda J_0 - I],$$

where

(a) J_r is a real Jordan form for $(I + \lambda_0 K_r)K_r^{-1}$ and has real spectrum;
(b) J_c is a real Jordan form for $(I + \lambda_0 K_c)K_c^{-1}$ with spectrum in $\mathbb{C}\backslash\mathbb{R}$;
(c) J_0 is a real Jordan form for $K_0(I + \lambda_0 K_0)^{-1}$ so that $\sigma(J_0) = \{0\}$. ∎

Notice that since $Y^{-1}AX = \operatorname{diag}[I, I, J_0]$, $Y^{-1}BX = \operatorname{diag}[J_r, J_c, I]$, and J_c is necessarily nonsingular, it follows that J_r and B have the same nullity, and J_0 and A have the same nullity.

Now we can prove an analogue of Theorem 4.1 for real pencils.

THEOREM 5.3. *If $\lambda A - B$ is a real regular H-selfadjoint pencil and H is real, then there exist real nonsingular matrices X and Y such that*

$$Y^{-1}(\lambda A - B)X = \begin{bmatrix} \lambda I - J & 0 \\ 0 & \lambda K - I \end{bmatrix}, \quad Y^T H Y = \begin{bmatrix} P_{\varepsilon_1, J} & 0 \\ 0 & P_{\varepsilon_2, K} \end{bmatrix}, \tag{5.2}$$

where $(P_{\varepsilon_1,J}, J)$, $(P_{\varepsilon_2,K}, K)$ are real selfadjoint canonical pairs and K is nilpotent.

Proof. According to Proposition 5.2 there exist nonsingular real matrices X_0 and Y_0 such that the first of equations (5.2) holds with J, K real Jordan matrices and K nilpotent. By Proposition 1.2 this pencil is $Y_0^T H Y_0$-selfadjoint. However, Theorem 2.4 applies and shows that $Y_0^T H Y_0 = \operatorname{diag}[H_1, H_2]$, $\lambda I - J$ is H_1-selfadjoint, $\lambda K - I$ is H_2-selfadjoint, and H_1, H_2 are real. Making the real canonical reductions of (H_1, J) and (H_2, K) we obtain the result. ∎

6. Canonical forms for real H-unitary pencils. It is clear that if A is H-unitary and also real, then the spectrum of A (and also the structure of a Jordan form for A) is symmetric with respect to both the real line and the unit circle. However, symmetry of this kind does not ensure that A is H-unitary for a real H. For example, if $A = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$, then $\sigma(A)$ has the necessary symmetry, but it is easily verified that there is no real Hermitian nonsingular H such that $A^T H A = H$ (although there is a nonsingular H with this property). We will return to the characterization of real matrices which are H-unitary for a real H in section 7.

A canonical form real H-unitary pencils (with H real) can be obtained from the next theorem.

THEOREM 6.1. *Let $\lambda L - M$ be a real regular H-unitary pencil (with H real). Then there exist real nonsingular matrices X and Y such that*

$$Y^{-1}(\lambda L - M)X = \begin{bmatrix} \lambda I - S & 0 & 0 \\ 0 & \lambda I_m - N & 0 \\ 0 & 0 & \lambda N^T - I_m \end{bmatrix}, \quad Y^T H Y = \begin{bmatrix} G & 0 & 0 \\ 0 & 0 & I_m \\ 0 & I_m & 0 \end{bmatrix}, \tag{6.1}$$

where N is a nilpotent Jordan matrix, m is the algebraic multiplicity of the zero eigenvalue of $\lambda L - M$ (assuming $m \neq 0$), G is real, symmetric, and nonsingular, and S is a real G-unitary matrix.

Proof. Using Proposition 5.2 we deduce that there exist real nonsingular matrices X_0 and Y_0 such that

$$Y_0^{-1}(\lambda L - M)X_0 = \operatorname{diag}[\lambda I - S, \lambda N_1 - I, \lambda I - N_2],$$

where S is nonsingular and N_1, N_2 are nilpotent. However, Proposition 3.3(b) applies and shows that N_1 and N_2 are similar. Thus, by applying appropriate real similarities we find real nonsingular X_1 and Y_1 such that

$$Y_1^{-1}(\lambda L - M)X_1 = \operatorname{diag}[\lambda I - S, \lambda N^T - I_m, \lambda I_m - N], \tag{6.2}$$

where N is nilpotent and, (by Proposition 1.2), this pencil is unitary with respect to the real matrix $Y_1^T H Y_1$. Using this unitary property it is easily seen that

$$Y_1^T H Y_1 = \begin{bmatrix} G & 0 & 0 \\ 0 & 0 & K^T \\ 0 & K & 0 \end{bmatrix}$$

for some real nonsingular K which commutes with N^T, and S is G-unitary.

Now define the real nonsingular matrix

$$T = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & K^{-1} \\ 0 & I_m & 0 \end{bmatrix}$$

and, using the fact that $KN^TK^{-1} = N^T$, it is found that

$$T^{-1}Y_1^{-1}(\lambda A - B)X_1 T = \operatorname{diag}[\lambda I - S, \lambda I_m - N, \lambda N^T - I_m],$$

and this pencil is unitary with respect to

$$T^T(Y_1^T H Y_1)T = \begin{bmatrix} G & 0 & 0 \\ 0 & 0 & I_m \\ 0 & I_m & 0 \end{bmatrix}. \qquad \square$$

7. Real H-unitary matrices. A question of independent interest which arose in section 6 concerns the characterization of those real matrices A which are H-unitary for some real nonsingular H, i.e., for which $A^THA = H$. It has been observed that the spectrum and Jordan structure of A are necessarily symmetric with respect to both the real line and the unit circle. In particular, eigenvalues can occur in the following combinations:

1. Eigenvalues at $+1$ or -1.
2. Conjugate pairs of nonreal unimodular eigenvalues.
3. Pairs of real eigenvalues λ and λ^{-1} ($\lambda \neq 0$, $\lambda \neq \pm 1$).
4. Quadruples of nonreal, nonunimodular eigenvalues $\alpha, \overline{\alpha}, \overline{\alpha}^{-1}, \alpha^{-1}$.

THEOREM 7.1. *A real square matrix A admits a real, symmetric, nonsingular solution H of $A^THA = H$ if and only if the spectrum of A has the necessary symmetry properties mentioned above and, in addition, the total number of Jordan blocks with eigenvalues $+1$ and even size is even, together with a similar property for the eigenvalue -1.*

We establish the theorem via a sequence of lemmas. First, by combining the arguments from the proof of Theorem 5.3 with the H-unitary complex case one can get the following for A and H from Theorem 6.1.

LEMMA 7.2. *Let A be an H-unitary matrix. Then*

(a) *there exists a real nonsingular T such that $T^{-1}AT$ is block diagonal with the blocks corresponding to different subsets of the spectrum of A of types 1 to 4 above;*

(b) *$T^{-1}AT$ is T^THT unitary and T^THT has the same block structure as $T^{-1}AT$.*

This lemma allows us to investigate each of the subsets 1 to 4 of the spectrum of A independently of the others.

Our attention is focused mainly on the question, Which real matrices admit (symmetric) nonsingular real solutions H of $A^THA = H$? We are also going to discuss

the set of all such solutions. It turns out that, in some sense, the first question depends on the second. We will say that *a real matrix A is of class $U(\mathbb{R})$* if A admits a nonsingular real symmetric solution H of $A^THA = H$. A nonsingular real symmetric solution will be said to be *proper*.

Notice that the matrix $A = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ mentioned above has just one Jordan block of even size with eigenvalue $+1$ and is not in $U(\mathbb{R})$. However, the direct sum of two such blocks *is* in $U(\mathbb{R})$. These examples suggest that at least when case 1 applies, canonical forms for real H-unitary matrices cannot be based only on the Jordan structure of A. Some additional structure is required. This issue is postponed and the next lemma concerns cases 3 and 4 above.

LEMMA 7.3. *Let $A \in U(\mathbb{R})$ and suppose, in addition, that the eigenvalues of A are nonunimodular. A real matrix G is a proper solution of $A^TGA = G$ if and only if there exists a real nonsingular matrix Q such that*

$$Q^{-1}AQ = \begin{bmatrix} K_0^{T^{-1}} & 0 \\ 0 & K_0 \end{bmatrix}, \qquad Q^TGQ = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}, \tag{7.1}$$

where K_0 is a real matrix corresponding to the spectral data of A in $|z| < 1$.

Proof. One direction immediately follows from (7.1). Conversely, given $A^TGA = G$ with G real, symmetric, and nonsingular, and the fact that there are no unimodular eigenvalues, it follows that the real invariant subspaces of A associated with the interior and exterior of the unit circle are G-neutral. Also, since the spectrum of A is symmetric about the real line, there is a real nonsingular S such that

$$A_0 = S^{-1}AS = \begin{bmatrix} K_0^T & 0 \\ 0 & K_0^{T^{-1}} \end{bmatrix},$$

and

$$G_0 = S^TGS = \begin{bmatrix} 0 & G_1 \\ G_1^T & 0 \end{bmatrix}.$$

The equality $A_0^TG_0A_0 = G_0$ gives us $G_1^TK_0^T = K_0G_1^T$. Define now

$$T = \begin{bmatrix} 0 & G_1^{T^{-1}} \\ I & 0 \end{bmatrix}.$$

Using $G_1^TK_0^T = K_0G_1^T$ it is found that

$$T^TG_0T = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix}, \qquad T^{-1}A_0T = \begin{bmatrix} K_0^{T^{-1}} & 0 \\ 0 & K_0 \end{bmatrix}. \qquad \square$$

We treat eigenvalues of case 2 in a sequence of lemmas culminating in Corollary 7.7.

Real matrices of the form $S = \left[\begin{smallmatrix} a & b \\ -b & a \end{smallmatrix}\right]$ appear in real Jordan forms when case 2 applies and it will be convenient to denote the commutative algebra of all such matrices by $\mathcal{S}^{2\times 2}$. As part of the "canonical real Jordan" structure used in Proposition 5.1, we know that any real matrix having only nonreal unimodular spectrum is similar to

a direct sum of blocks of the following kind:

$$\begin{bmatrix} F & I & 0 & \dots & 0 \\ 0 & F & I & & \vdots \\ \ddots & & & \ddots & \\ & & & \ddots & I \\ 0 & & \dots & & F \end{bmatrix}, \tag{7.2}$$

where all entries are real 2×2 matrices, $F \in \mathcal{S}^{2\times 2}$, $a^2 + b^2 = 1$, and $b \neq 0$.

PROPOSITION 7.4. *If $F \in \mathcal{S}^{2\times 2}$, $a^2 + b^2 = 1$, $b \neq 0$, and $B \in \mathbb{R}^{2\times 2}$, then the following are equivalent:*

(i) $F^T BF - B \in \mathcal{S}^{2\times 2}$,
(ii) $FB = BF$,
(iii) $B \in \mathcal{S}^{2\times 2}$.

Proof. Note that under the given hypotheses, $F^T = F^{-1}$, and the equivalence of (i) and (ii) follows immediately. Then observe that if we write $B = \left[\begin{smallmatrix} x & y \\ z & w \end{smallmatrix}\right]$, then

$$FB - BF = b \begin{bmatrix} y + z & w - x \\ w - x & -(y + z) \end{bmatrix},$$

with $b \neq 0$. From the definition of $\mathcal{S}^{2\times 2}$ it follows immediately that $FB - BF \in \mathcal{S}^{2\times 2}$ if and only if $y + z = 0$ and $w - x = 0$. But these statements are equivalent to either $FB - BF = 0$, or $B \in \mathcal{S}^{2\times 2}$. ∎

LEMMA 7.5. *Let $A_1 \in \mathbb{R}^{2k_1\times 2k_1}$, $A_2 \in \mathbb{R}^{2k_2\times 2k_2}$ ($k_1 \leq k_2$; $k_1, k_2 \in \mathbb{N}$) have the form* (7.2). *If for a real matrix K ($\in \mathbb{R}^{2k_1\times 2k_2}$) the equation*

$$A_1^T K A_2 = K \tag{7.3}$$

holds, then K (being divided into 2×2 blocks ($K = (K_{ij})_{i=1}^{k_1}{}_{j=1}^{k_2}$; $K_{ij} \in \mathbb{R}^{2\times 2}$)) has the following block-triangular form:

(a) $K_{ij} \in \mathcal{S}^{2\times 2}$,
(b) $K_{ij} = 0$ *for $j < k_2 - i$ (if any).*

Proof. Direct computations give us $F^T K_{11} F - K_{11} = 0$, so, by Proposition 7.4, $K_{11} \in \mathcal{S}^{2\times 2}$. In the first block row we get $F^T K_{1,i-1} + F^T K_{1i} F = K_{1i}$ for $i \geq 2$. Then, since $K_{1,i-1} \in \mathcal{S}^{2\times 2}$, proceeding step-by-step, we get the required form for K_{1i}. Moreover, $K_{1,i-1} = FK_{1i} - K_{1i}F$ for $2 \leq i \leq k_2$, so we also have $K_{1,i-1} = 0$. Thus, $K_{11} = \cdots = K_{1,k_2-1} = 0$, and K_{1,k_2} has the required form.

In the same way we find that in the first block column $K_{11} = \cdots = K_{k_1-1,1} = 0$ (and the required form for $K_{k_1,1}$).

For the rest of the blocks (if any)

$$K_{i-1,j-1} + F^T K_{i,j-1} + K_{i-1,j} F + F^T K_{ij} F = K_{ij}.$$

Thus, first of all, all blocks have the required form and, moreover, have $k_2 - 1$ zeros in the first block row and $k_1 - 1$ zeros in the first block column. It follows that the ith block row has one zero block less than in the block row with number $i - 1$. ∎

LEMMA 7.6. *If $A \in \mathbb{R}^{2k\times 2k}$ and has the form* (7.2), *then*

(a) $A \in U(\mathbb{R})$.
(b) *All real nonsingular symmetric solutions H of $A^T HA = H$ are triangular as described in Lemma* 7.5, *and the 2×2 blocks $A_{i,k-i+1}$ are (up to the same multiplicative constant) orthogonal 2×2 matrices.*

Proof. We start with $k = 1, 2$. These cases can be checked by direct computations:

If $k = 1$, then $A = F$, which is an orthogonal matrix.

If $k = 2$, $A = \left[\begin{smallmatrix} F & 1 \\ 0 & F \end{smallmatrix}\right]$ is $\left[\begin{smallmatrix} 0 & H_1 \\ H_1^T & H_2 \end{smallmatrix}\right]$-unitary, where H_1 is a real 2×2 orthogonal matrix corresponding to rotation through $\pi/2 - \alpha$, where α defines the rotation of F, and H_2 is a scalar matrix. These two cases provide all proper solutions of $A^T H A = H$ (up to multiplicative constants).

We now combine inductive arguments with the previous lemma in the following way.

Let A_0 be a $2k \times 2k$ matrix of the form (7.2). Define now

$$A = \begin{pmatrix} I & I & \dots & 0 \\ 0 & & & \vdots \\ \vdots & & A_0 & I \\ 0 & \dots & 0 & I \end{pmatrix}; \tag{7.4}$$

we are looking for a proper solution of $A^T H A = H$ in the following form:

$$H = \begin{pmatrix} 0 & \dots & 0 & G_0 \\ & & & G_1 \\ \vdots & & H_0 & \vdots \\ 0 & & & G_k \\ G_0^T & G_1^T \dots & G_K^T & S \end{pmatrix}, \tag{7.5}$$

where H_0 is a proper solution of the $2k \times 2k$ equation $A_0^T H_0 A_0 = H_0$. In such a situation it is enough to define the last block column of H. In other words one should equate the blocks of the last block columns of H and $A^T H A$. Let us denote the blocks of the last block column of H_0 by $L_1, \dots, L_k$.

The equation $H = A^T H A$ yields for the blocks mentioned above

$$F^T G_0 F = G_0, \quad L_{i-1} + F^T L_i + G_{i-1} F + F^T G_i F = G_i \quad (i = 1, \dots, k, L_0 = 0),$$

$$F^T G_k^T + L_k + G_k F + F^T S F = S.$$

According to Lemma 7.5(a) all the blocks commute, so one can rewrite the equalities in the following way:

$$G_0 = G_0, \quad F^T L_1 + G_0 F = 0, \quad L_{i-1} + F^T L_i + G_{i-1} F = 0 \quad (i = 2, \dots, k),$$

$$F^T G_k^T = L_k + G_k F = 0.$$

Finally,

$$G_0 = -L_1 F^{-2}, \quad G_1 = -L_2 F^{-2} - L_1 F^{-1}, \dots, G_{k-1} = -L_k F^{-2} - L_{k-1} F^{-1}.$$

The last equality $G_k F + (G_k F)^T + L_k = 0$ requires some explanation. We see at this step that the block S is not involved in any equation that has been considered. According to general conditions, it must be a scalar matrix. For H_0 the role of S is played by L_k. It is also a scalar matrix. The equation $G_k F + (G_k F)^T + L_k = 0$ will always possess solutions of the required form.

Using $G_0 = -L_1F^{-1}$ and (7.5), H is found to be nonsingular. All 2×2 blocks $A_{i,k-i+1}$ are (up to the same multiplicative constant) defined as pairs of adjoint orthogonal matrices. In the odd case the central element is I. ∎

COROLLARY 7.7. *Any real matrix A for which all eigenvalues are nonreal and unimodular is in $U(\mathbb{R})$. (Indeed, one can define H as a direct sum of blocks corresponding to a block diagonal real Jordan form of A).*

Now we pass to the remaining case: Describe the possible Jordan structures of matrix $A = I + N$, where N is nilpotent and when $A \in U(\mathbb{R})$. (The cases with eigenvalues -1 or $+1$ and -1 are easily included.) According to an example mentioned above, not every matrix $I + N$ is in $U(\mathbb{R})$, although the direct sum of two such blocks may be in $U(\mathbb{R})$.

We use the following notation for joint representation of ℓ Jordan chains of length k (it is permutationally similar to a direct sum of l Jordan blocks of size k):

$$J_{k,l} := \begin{bmatrix} I & I & \dots & 0 \\ & & \ddots & \\ & & \ddots & I \\ 0 & \dots & & I \end{bmatrix}, \qquad I \in \mathbb{R}^{l\times l}, \quad J_{k,l} \in \mathbb{R}^{kl\times kl}.$$

Using the arguments and the methods of Lemma 7.6 and Lemma 7.5(b) one can prove the following lemma.

LEMMA 7.8.

(a) *Suppose that $k_1 \le k_2$ and for some real matrix K*

$$J_{k_1,l_1}^T K J_{k_2,l_2} = K.$$

Then K (being partitioned into $l_1 \times l_2$ blocks) has the following triangular form: $K_{ij} = 0$ for $j \le k_2 - i$.

(b) *$J_{k,l}$ is in $U(\mathbb{R})$ if and only if $J_{k+2,l}$ is in $U(\mathbb{R})$.*

Remark 7.9. In the case $k_1 = k_2$, $l_1 = l_2$, $K = K^T$ block elements on the diagonal $\{K_{i,k-i+1}\}_{i=1}^k$ have the same rank.

LEMMA 7.10.

(a) *$J_{2p-1,l}$ is in $U(\mathbb{R})$, $(l, p \in \mathbb{N})$.*

(b) *$J_{2p,l}$ is in $U(\mathbb{R})$, $(l, p \in \mathbb{N})$ if and only if l is even.*

Proof. According to Lemma 7.8 one can decide whether J_{kl} is in $U(\mathbb{R})$ by considering only $J_{1,l}$ and $J_{2,l}$.

Obviously, $J_{1,l}$ is in $U(\mathbb{R})$. For $J_{2,l}$ one can write

$$\begin{pmatrix} I & 0 \\ I & I \end{pmatrix} \begin{pmatrix} 0 & H_1 \\ H_1^T & H_2 \end{pmatrix} \begin{pmatrix} I & I \\ 0 & I \end{pmatrix} = \begin{pmatrix} 0 & H_1 \\ H_1^T & H_2 \end{pmatrix}, \qquad 0, I, H_i \in \mathbb{R}^{l\times l}.$$

This equality holds if and only if $H_1^T + H_1 = 0$. It is also required that $\left(\begin{smallmatrix} 0 & H_1 \\ H_1^T & H_2 \end{smallmatrix}\right)$ must be invertible, thus $\det H_1 \neq 0$. In this case one can find a nonsingular solution of $H_1^T + H_1 = 0$ if and only if the size of H_1 is even. (When the size is odd, $H_1^T + H_1 = 0$ implies $\det H_1 = -\det H_1$.) ∎

Now we are ready to take the final step. Any matrix $A = I + N$ where N is nilpotent has a unique representation by a similar matrix of the following form:

$$\text{Diag}\,(J_{k_i,l_i})_{i=1}^s \text{ with } k_1 < \dots < k_s. \tag{7.6}$$

The following lemma shows that a representation of A with the form (7.6) is in $U(\mathbb{R})$ if and only if every matrix J_{k_i,l_i} $(i = 1, \dots, s)$ is in $U(\mathbb{R})$. An application of Lemma 7.10 then completes the proof of Theorem 7.1.

LEMMA 7.11. *Let A be in the form* (7.6) *and H be a nonsingular real symmetric solution of $A^T HA = H$. Then submatrices of H corresponding to the diagonal blocks of A are invertible.*

Proof. Let us denote by $H^{(1)}, \dots, H^{(s)}$ the diagonal submatrices of H in the positions of $J_{k_1,l_1}, \dots, J_{k_s,l_s}$ in A. Note that $H^{(r)} = (H^{(r)})^T$, $J^T_{k_r,l_r} H^{(r)} J_{k_r,l_r} = H^{(r)}$ $(r = 1, \dots, s)$.

All $H^{(r)}$ are block triangular (Lemma 7.8) when partitioned into $l_r \times l_r$ blocks. According to Remark 7.9, $H^{(r)}$ is invertible if and only if the last block element (we denote it by $H^{(r)}_{1,k_r}$ $(\in \mathbb{R}^{l_r \times l_r})$) of the first block row has full rank. We are going to show that all such blocks $H^{(1)}_{1,k_1}, \dots H^{(s)}_{1,k_s}$ are invertible, and this will complete the proof.

Let A be in the form (7.6). We start with $H^{(s)}$ (i.e., the submatrix of H corresponding to the collection of all chains of A of maximal length; in other words, $H^{(s)}$ (or J_{k_s,l_s}, resp.) is the last diagonal block of H (of A, resp.)).

According to Lemma 7.8 all elements of H placed to the left of $H^{(s)}_{1,k_s}$ are equal to zero. Since H is invertible, $H^{(s)}_{1,k_s}$ is also nonsingular.

Note also that since $H = H^T$, the lowest block element of the first block column of $H^{(s)}$ equals $(H^{(s)}_{1,k_s})^T$ (according to our notation it is $H^{(s)}_{k_s,1}$) and all the elements of H above it (above $H^{(s)}_{k_s,1}$) are zeros.

Thus, according to a rule for developing determinants, one finds that the matrix obtained from H by omitting all rows and columns which intersect at least one of $H^{(s)}_{1,k_s}$ or $H^{(s)}_{k_s,1}$ is also nonsingular. The new matrix $\hat{H}$ must not be precisely of the form corresponding to (7.6), but now (Lemma 7.8) all elements of $\hat{H}$ placed in the rows of $H^{(s-1)}_{1,k_s-1}$ (the last block element of the first block row of $H^{(s-1)}$) which do not belong to $H^{(s-1)}_{1,k_s-1}$ are zeros. This allows us to apply the algorithm used above in order to show that $H^{(s-1)}_{1,k_s-1}$ is nonsingular; $H^{(s-1)}$ is also nonsingular.

Using this algorithm s times, one obtains the invertibility of all submatrices $H^{(1)}, \dots, H^{(s)}$. □

Examination of the main diagonal blocks in the relation $A^T HA = H$ now shows that each matrix J_{k_i,l_i} is in $U(\mathbb{R})$.

Appendix. In this appendix we summarize some known results for regular pencils $\lambda A - B$ with no symmetry assumptions on A and B. We begin with useful results which are carefully developed in reference [5]; see also [11] and [12] for earlier treatments. These sources present the results in the context of Banach spaces. We specialize to the case in which A and B are $n \times n$ matrices. As in the main body of this paper, $\lambda A - B$ is always a regular pencil with $A, B \in \mathbb{C}^{n \times n}$, generally both singular.

The symbol Γ will denote a closed Cauchy contour in $\mathbb{C}$ (see [5] for details), and Δ_+, Δ_- denote the inner domain and outer domain of Γ (including ∞), respectively. Generalized Riesz projections are defined in the first lemma.

LEMMA A.1. *Let Γ be a Cauchy contour with $\Gamma \cap \sigma(\lambda A - B) = \emptyset$, and*

$$P = \frac{1}{2\pi i} \int_\Gamma (\lambda A - B)^{-1} A \, d\lambda, \qquad Q = \frac{1}{2\pi i} \int_\Gamma A(\lambda A - B)^{-1} \, d\lambda. \tag{A.1}$$

Then P, Q are projections onto the spectral deflating subspaces of $\lambda A - B$ and $\lambda A^T - B^T$, respectively, associated with $\Delta_+ \cap \sigma(\lambda A - B)$.

THEOREM A.2. *Let P, Q be the projections of* (A.1) *and let $X = [X_1 \;\; X_2]$, $Y = [Y_1 \;\; Y_2]$ be nonsingular matrices for which*

$$\operatorname{Im} X_1 = \operatorname{Im} P, \quad \operatorname{Im} X_2 = \operatorname{Ker} P, \quad \operatorname{Im} Y_1 = \operatorname{Im} Q, \quad \operatorname{Im} Y_2 = \operatorname{Ker} Q.$$

Then

$$Y^{-1}(\lambda A - B)X = \begin{bmatrix} \lambda A_1 - B_1 & 0 \\ 0 & \lambda A_2 - B_2 \end{bmatrix}, \tag{A.2}$$

where A_1 is invertible, $\lambda A_2 - B_2$ is regular, and

$$\sigma(\lambda A_1 - B_1) = \Delta_+ \cap \sigma(\lambda A - B), \quad \sigma(\lambda A_2 - B_2) = \Delta_- \cap \sigma(\lambda A - B). \tag{A.3}$$

Furthermore, if $0 \in \Delta_+$, then B_2 is invertible and if $\Delta_- = \{\infty\}$, then A_2 is nilpotent.

Note that if r is the dimension of $\operatorname{Im} P$ (and $\operatorname{Im} Q$), then $\lambda A_1 - B_1$ is $r \times r$. In the body of the paper, we use deflations like the right side of (A.2) in which $A_1 = I_r$ and $B_2 = I_{n-r}$. For discussion of numerically stable methods for the computation of deflations like (A.2) the reader is referred to [2].

PROPOSITION A.3.

(a) *If $0 \in \Delta_+$, then there exist nonsingular X and Y such that*

$$Y^{-1}(\lambda A - B)X = \begin{bmatrix} \lambda I_r - B_1 & 0 \\ 0 & \lambda A_2 - I_{n-r} \end{bmatrix} \tag{A.4}$$

and $\sigma(\lambda I_r - B_1) \subset \Delta_+$, $\sigma(\lambda A_2 - I_{n-r}) \subset \Delta_-$.

(b) *When* (A.4) *holds with nonsingular X and Y and we write $X = [X_1 \;\; X_2]$, $Y = [Y_1 \;\; Y_2]$, where X_1, Y_1 are $n \times r$, then $\operatorname{Im} X_1$ and $\operatorname{Im} X_2$ are deflating subspaces of $\lambda A - B$ associated with $\sigma(\lambda I_r - B_1)$ and $\sigma(\lambda A_2 - I_{n-r})$, respectively, and*

$$Y_1 = AX_1, \qquad Y_2 = BX_2. \tag{A.5}$$

Proof. Part (a) follows immediately from Theorem A.2. For (b) observe that (A.4) implies

$$AX_1 = Y_1, \quad AX_2 = Y_2A_2, \quad BX_1 = Y_1B_1, \quad BX_2 = Y_2 \tag{A.6}$$

and the conclusions follow from these relations (see Lemma 1.6.1 and what follows in [9], for example). ∎

We will also take advantage of the following result (see [10] and Theorem IV.2.1 of [5]).

THEOREM A.4. *Let A_1, B_1 be $r \times r$ matrices, A_2, B_2 be $(n-r) \times (n-r)$, and C be $r \times (n-r)$. If the spectra of $\lambda A_1 - B_1$ and $\lambda A_2 - B_2$ do not intersect, then the equation*

$$B_1ZA_2 - A_1ZB_2 = C$$

has a unique solution Z.

Acknowledgments. The authors are grateful to L. Rodman and anonymous reviewers for comments which have helped to improve the exposition of this paper.

REFERENCES

[1] D. J. Clements and K. Glover, *Spectral factorization via hermitian pencils*, Linear Algebra Appl., 122/123 (1989), pp. 797–846.
[2] J. W. Demmel and B. Kagstrom, *Computing stable eigendecomposition of matrix pencils*, Linear Algebra Appl., 88/89 (1987), pp. 139–186.
[3] L. Elsner and P. Lancaster, *The spectral variation of pencils of matrices*, J. Comput. Math., 3 (1985), pp. 262–274.
[4] J. D. Gardiner and A. J. Laub, *A generalization of the matrix-sign-function solution for algebraic Riccati equations*, Internat. J. Control, 44 (1986), pp. 823–832.
[5] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators*, Vol. 1, Birkhäuser-Verlag, Basel, Switzerland, 1990.
[6] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Academic Press, New York, 1982.
[7] I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Birkhäuser-Verlag, Basel, Switzerland, 1983.
[8] I. Krupnik and P. Lancaster, *Minimal pencil realizations of rational matrix functions with symmetries*, Canad. Math. Bull., to appear.
[9] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, Oxford, 1995.
[10] G. W. Stewart, *On the sensitivity of the eigenvalue problem* $Ax = \lambda Bx$, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.
[11] F. Stummel, *Diskreterkonvergenz linearer Operatoren* II, Math. Zeitschrift, 120 (1971), pp. 231–264.
[12] P. Van Dooren, *Factorization of a rational matrix: The singular case*, Integral Equations Operator Theory, 7 (1984), pp. 704–741.
[13] H. K. Wimmer, *Normal forms of symplectic pencils and the discrete time algebraic Riccati equation*, Linear Algebra Appl., 147 (1991), pp. 411–440.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 325–354, April 1998

003

APPLICATIONS OF THE DULMAGE–MENDELSOHN DECOMPOSITION AND NETWORK FLOW TO GRAPH BISECTION IMPROVEMENT*

CLEVE ASHCRAFT† AND JOSEPH W. H. LIU‡

Abstract. In this paper, we consider the use of the Dulmage–Mendelsohn decomposition and network flow on bipartite graphs to improve a graph bisection partition. Given a graph partition $[S, B, W]$ with a vertex separator S and two disconnected components B and W, different strategies are considered based on the Dulmage–Mendelsohn decomposition to reduce the separator size $|S|$ and/or the imbalance between B and W. For the case when the vertices are weighted, we relate this to the bipartite network flow problem. A further enhancement to improve a partition is to generalize the bipartite network to a general network and then solve a max-flow problem. We demonstrate the utility of these improvement techniques on a set of sparse test matrices, where we find top-level separators, nested dissection, and multisection orderings.

Key words. Dulmage–Mendelsohn decomposition, network flow, graph bisection, ordering algorithms, nested dissection, multisection

AMS subject classifications. 65F05, 65F50, 68R10

PII. S0895479896308433

1. Introduction.

The ability to find a good separator for a graph is necessary in many application areas [16], [29]. Our motivation to consider this problem is to determine good sparse matrix orderings for direct factorization methods [4], [5], [19], [24].

In a recent paper [2], the authors have applied the notion of blocking to obtain an efficient graph partitioning scheme to find a good vertex separator. The approach has three basic steps. In the first step, we construct a *domain decomposition* of the graph, consisting of a subset of vertices (called a multisector) whose removal decomposes the graph into a number of *domains*. Each domain is a connected subset of vertices. The second step uses a variant of the Kernighan–Lin scheme [21] on the set of domains to determine an approximation to a good separator. The last step refines the separator using some techniques from bipartite graph matching. One purpose of this paper is to give a full explanation of the machinery used in the separator improvement step.

The fundamental tool used in this final step is the *Dulmage–Mendelsohn decomposition* [8], which is a canonical decomposition of a bipartite graph based on the notion of matching. This decomposition has been used extensively to extract a *vertex separator* from an *edge separator* [15], [20], [23], [27]. The vertices that are incident to an edge in the edge separator form a *wide* vertex separator. A vertex separator is a *cover* for the edge separator if all edges are incident to a vertex of the separator. Using the Dulmage–Mendelsohn decomposition, one can find one or more vertex covering separators of minimum size that are subsets of this wide vertex separator.

*Received by the editors August 23, 1996; accepted for publication (in revised form) by Z. Strakoš January 30, 1997.

http://www.siam.org/journals/simax/19-2/30843.html

†Boeing Information and Support Services, P. O. Box 24346, Mail Stop 7L-22, Seattle, WA 98124 (cleve@espresso.rt.cs.boeing.com). This research was supported in part by ARPA contract DABT63-95-C-0122.

‡Department of Computer Science, York University, North York, Ontario, Canada M3J 1P3 (joseph@cs.yorku.ca). This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant A5509 and in part by ARPA contract DABT63-95-C-0122.

This decomposition has been used to improve a vertex separator in earlier papers [24], [25], [26]. Let the vertices in the graph be partitioned as a vertex separator S and two components B and W. We consider the edge separator that contains edges linking vertices in S to B (or S to W). This defines a wide vertex set containing vertices in S and all vertices in B (or W) adjacent to S. We use the Dulmage–Mendelsohn decomposition to find a covering separator of minimum size from this set of vertices[1].

The same technique can be applied to the new separator and its new adjacent sets, so the overall improvement process is iterative in nature. At each step, a wide vertex separator is taken from the current separator and one of the two components, and a covering vertex separator subset of minimum size is obtained. It is accepted as the new vertex separator only if the quality of its induced partition is better.

In this paper, we consider related approaches, initially developed in [2], to improve a vertex separator. Although the Dulmage–Mendelsohn decomposition is defined only for unit-weight graphs, we are able to extend it to a special class of weighted graphs, thus greatly reducing the execution time in many cases. In the extension, we reformulate it into a much-studied combinatorial problem involving the flow of commodities through an interconnected network: a maximum network flow problem [9], [11], [12], [18], [28]. The solution to our separator improvement step is thus transformed to solving a maximum flow problem on a bipartite network. We also relate the improved separator in the new partition to the min-cut set in the well-known max-flow min-cut theorem on network flows.

We have explored an additional advantage in the transformation of a bipartite graph matching problem to a bipartite network flow problem. By adding and deleting edges from a bipartite network it may be possible to construct a new network that yields a smaller separator. The new network would not be bipartite and the new separator need not be a covering separator. By adding vertices and edges we can generate larger networks that might yield still smaller separators.

An outline of this paper is as follows. In section 2, we give a formal description of the partition improvement problem and introduce the various notations used throughout the paper. Section 3 starts with a discussion on reducing the size of a separator using bipartite graph matching. This provides the motivation for using the Dulmage–Mendelsohn decomposition for bipartite graphs. This section is mainly expository in nature; the results can be found in [24], [25], and [26]. Section 4 considers the use of the Dulmage–Mendelsohn decomposition to improve the balance of a partition.

In section 5, we introduce the notion of a *compressed* graph induced by a grouping of vertices that share the same adjacent sets. Compressed graphs can be considered as a special kind of weighted graphs. The Dulmage–Mendelsohn decomposition is then generalized to handle compressed bipartite graphs.

In section 6, we relate this decomposition of a compressed bipartite graph to a max-flow solution to a bipartite network problem. We point out the equivalence between the generalized matching and a max-flow, and between the improved separator and a min-cut. We also describe a new enhancement where we transform the bipartite network into a larger, more general network based on the underlying graph structure. We show that a max-flow min-cut solution to this new general network is at least as good as and is often better than that of the bipartite network. We also generalize the network flow approach to even wider separators formed from the separator and many "layers" of adjacent sets from one or both components in the partition.

[1] It is a little-appreciated fact that a covering separator of minimum size may *not* be a separator of minimum size, i.e., a separator of minimum size may not be incident on all the edges of the edge separator. (See the example graph in Figures 2 and 8.)

Section 7 contains experimental results on separator/partition improvements. We compare the improvement in partitions based on solving a max-flow problem on a bipartite network, the induced two-layer network, and a centered three-layer wide network. Sparse matrix ordering statistics are also given when these techniques are used in a nested dissection and a multisection ordering code [3]. The multisection statistics are at least as good and often are better than those from the multiple minimum degree ordering approach. Section 8 contains our concluding remarks.

2. Definitions and notations. Let $G = (V, E)$ be a given undirected graph. The adjacent set of a vertex v is given by

$$Adj(v) = \{u \neq v \mid (u, v) \in E\}.$$

Without loss of generality, we assume the graph is connected. A *walk* is a sequence of vertices $v_0, v_1, \ldots, v_m$ such that $(v_i, v_{i+1}) \in E$. A *path* is a walk without any repeated vertices.

A vertex subset S is a *vertex separator* if the subgraph induced by the vertices in V but not in S has more than one connected component. An *edge separator* is a set of edges whose removal disconnects the graph. A separator is *minimal* if no subset of it forms a separator.

A *bisector* is a separator whose removal gives at least two connected components. We shall use the notation $[S, B, W]$ to represent a two-set partition, where the removal of the bisector S will give two disconnected portions B and W; that is, $Adj(B) \subseteq S$ and $Adj(W) \subseteq S$. We measure the *imbalance* of a partition as the dimensionless ratio $\max\{|B|, |W|\}/\min\{|B|, |W|\}$. We shall often assume that B is the bigger portion so that $|B| \geq |W|$ and the imbalance is $|B|/|W|$. Our objective is to determine a well-balanced partition with a small separator size $|S|$.

In this paper, we consider methods to improve a given partition. Therefore, we need to compare the quality of the original and the modified partitions. Following [2], we use this evaluation function

$$\gamma[S, B, W] = |S| \left(1 + \alpha \frac{\max\{|B|, |W|\}}{\min\{|B|, |W|\}}\right),$$

where α is some constant greater than zero. The separator size $|S|$ is the primary metric while the imbalance is used as a "penalty" multiplicative factor. A large value of the constant α places a large emphasis on the balance. We have used the penalty cost function $\gamma[S, B, W]$ with $\alpha = 1$ in all the experiments in section 7.

Throughout the paper we will be concerned with a subset of vertices, those vertices just "outside" the subset, and those "inside" the subset. To make these concepts clear we introduce the following notation. Let Y be a vertex subset of V. The *interior* of Y is defined to be

$$Int(Y) = \{y \in Y \mid Adj(y) \subseteq Y\},$$

and contains all nodes in Y that are adjacent to no nodes outside of Y. The *boundary* of Y, or its *adjacent set*, is the set of nodes not in Y that are adjacent to Y,

$$Adj(Y) = \{v \in V \setminus Y \mid (y, v) \in E \text{ for some } y \in Y\} = \left(\bigcup_{y \in Y} Adj(y)\right) \setminus Y.$$

```
PARTITION-IMPROVE [S, B, W]
    Improved = true
    while Improved do
        if |B| < |W| then interchange B and W // make B the larger portion
        if a subset Z of S is found with γ([S, B, W]_{Z↦W}) < γ[S, B, W] then
            [S, B, W] = [S, B, W]_{Z↦W}
        else
            if a subset Z of S is found with γ([S, B, W]_{Z↦B}) < γ[S, B, W] then
                [S, B, W] = [S, B, W]_{Z↦B}
            else
                Improved = false
            end if
        end if
    end while
```

FIG. 1. *Partition improvement scheme.*

The *border* of Y is a subset of Y, namely, the boundary of the interior of Y,

$$Border(Y) = Adj(Int(Y)) = Y \setminus Int(Y),$$

or those nodes in Y that are not in the interior of Y.

3. Partition improvement and the Dulmage–Mendelsohn decomposition.

3.1. A partition improvement algorithm by moves. Let $[S, B, W]$ be a two-set partition of a given graph G. Consider a subset Z of S. Let $Z \mapsto W$ be the move of Z to W that moves the subset Z from S to W, thereby creating the following new partition:

$$B_{Z\mapsto W} = B \setminus Adj(Z), \quad W_{Z\mapsto W} = W \cup Z, \quad \text{and} \quad S_{Z\mapsto W} = (S \setminus Z) \cup (Adj(Z) \cap B).$$

We use the notation $[S, B, W]_{Z\mapsto W}$ to refer to the new partition.

We consider a partition improvement scheme that uses moves by finding subsets Z that will help in reducing the evaluation function $\gamma[S, B, W]$. A high-level description of the improvement algorithm is described in Figure 1. The scheme makes a first attempt to reduce the evaluation function of the partition by moving a subset from S to the smaller portion W. If no such move can be found, it tries to improve the partition by moving a separator subset to the larger portion B. It continues until no reduction can be obtained.

3.2. Improving the separator size by graph matching. Recall that the evaluation function $\gamma[S, B, W]$ on a partition is given by the penalty function based on the separator size and the imbalance ratio. In practice, the weight α is chosen to be close to one so that the separator size has a strong influence on the partition evaluation. Therefore, one way to look for an improvement to the partition is to reduce the separator size. Consider the move $Z \mapsto W$. The new separator size is given by

$$|S_{Z\mapsto W}| = |S| - |Z| + |Adj(Z) \cap B|.$$

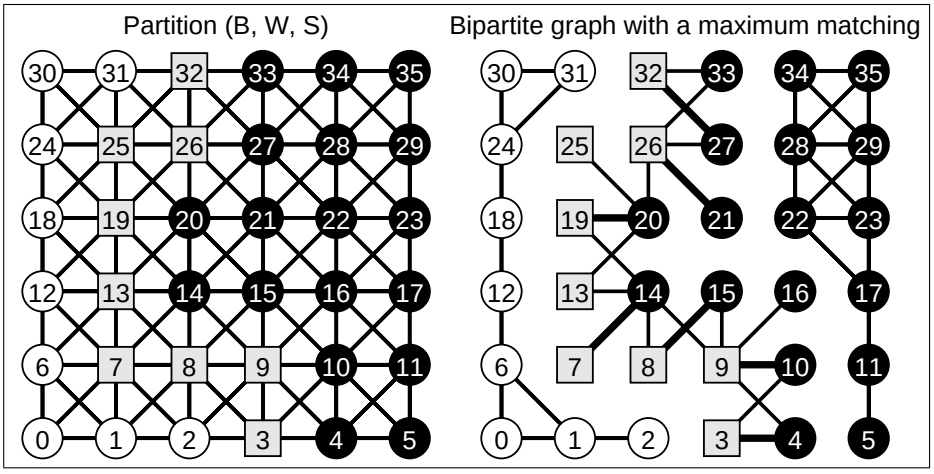


FIG. 2. *Bipartite graph example from a separator partition.*

Therefore, if we can find a subset Z of S such that $|Z| > |Adj(Z) \cap B|$, the move of Z to W will result in a reduction of the separator size by an amount of $|Z| - |Adj(Z) \cap B|$. (Note that this does not always guarantee a reduction in the evaluation function value.)

In [24], the technique of bipartite graph matching is used to find such a subset Z of S with $|Z| > |Adj(Z) \cap B|$. We shall first describe the necessary terminology in graph matching and state the results relevant to this approach.

A *bipartite graph* is an undirected graph whose node set can be divided into two disjoint sets X and Y such that every edge has one endpoint in X and the other in Y. A *matching* of a bipartite graph H is a subset M of edges such that no two edges in this subset have a node in common. A node that is incident to some edge in M is said to be *covered*; otherwise, it is *exposed*. If (x, y) belongs to the matching M, then $x = \text{mate}(y)$ and $y = \text{mate}(x)$. The number of edges in M is called the *size* of the matching. A *maximum matching* is one with the largest possible size. A *complete matching* is a matching of size $\min\{|X|, |Y|\}$.

We now consider the results in graph matching relevant to our context of improving a two-set partition $[S, B, W]$. Assume that B is the larger portion. Consider the bipartite graph $H = (S, Border(B), E_H)$, where E_H contains the set of edges between vertices in S and those in $Border(B)$ of the original graph G. Recall that $Border(B) = B \cap Adj(S)$. For simplicity, we often refer to this bipartite graph by $H(S, B)$, and the two defining sets as S and B. However, it is implicit that only the subset $Border(B)$ of B is used in H. For a node x in this bipartite graph H, we shall use $Adj_H(x)$ to represent the set of adjacent nodes of x in the bipartite graph H. We extend the notation to $Adj_H(U)$ for the adjacent set of a subset U of nodes. Note that we use $Adj(x)$ and $Adj(U)$ to represent the adjacent sets in the original graph G. It should be clear from the definition of H that for any subset Z of S, $Adj_H(Z) = Adj(Z) \cap B$.

In Figure 2, we illustrate the induced bipartite graph H for a 6×6 grid problem with the 9-point operator; that is, each interior node is connected to its eight neighbors. For the given separator of size 9, we obtain its associated bipartite graph H. In the figure, a matching between S and B is also given; and the edges in the matching are indicated by thick lines. This matching is of size 7 and it is maximum.

In the separator improvement scheme, we want to find a subset Z of S satisfying $|Z| > |Adj(Z) \cap B|$. The next theorem by Hall [14] relates the *nonexistence* of such a subset with bipartite graph matching.

THEOREM 3.1 (see Hall [14]). *The bipartite graph H has a complete matching of S into B if and only if for every subset Z of S, $|Z| \leq |Adj(Z) \cap B|$.*

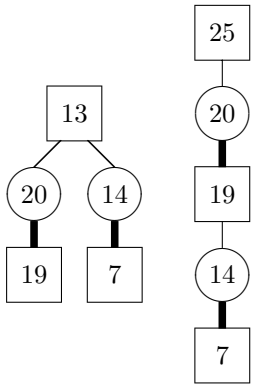
Theorem 3.1 can be used to provide a necessary and sufficient condition for the existence of a size-improving subset Z of S. The condition is that the bipartite graph H does not have a complete matching from S into B. This implies that for a maximum matching, there will be some exposed nodes in S, that is, nodes without a mate in the matching. In the example of Figure 2, there are two exposed nodes 13 and 25 in S so that the maximum matching is not complete. We know from Theorem 3.1 that we can find some size-improving subset Z of S.

To discuss the way to find such subsets, we need the notion of an alternating path. For a given matching M, consider a path $\langle x_0, x_1, ..., x_k \rangle$ where no vertex is repeated. It is called *alternating* with respect to M if the alternate edges belong to the matching M. For example, in Figure 2, the path $\langle 25, 20, 19, 14, 7 \rangle$ is alternating; the edges $(20, 19)$ and $(14, 7)$ belong to the matching. In [24], alternating paths are used in the following result to find a subset Z satisfying $|Z| > |Adj(Z) \cap B|$.

THEOREM 3.2 (see Liu [24]). *Let $x \in S$ be an exposed node in a maximum matching of H. Define $S_x = \{s \in S \mid s \text{ is reachable from } x \text{ via alternating paths}\}$. Then $|S_x| - |Adj_H(S_x)| = 1$.*

The set S_x can be determined by performing a special kind of breadth-first search starting from the exposed node x. The search is restricted to nodes reachable via alternating paths. Then S_x is given by the nodes of S appearing in this *breadth-first search tree* rooted at x. Since we only consider alternating paths in the traversal, we shall refer to this tree as an *alternating breadth-first search tree*. This set S_x can be used as Z to reduce the separator size by one.

For the example in Figure 2, there are two exposed separator nodes: 13 and 25. Immediately below we find the two alternating breadth-first search trees that start from 13 and 25, respectively. It is clear that $S_{13} = \{7, 13, 19\}$ and $Adj_H(S_{13}) = \{14, 20\}$. On the other hand, $S_{25} = \{7, 19, 25\}$ and $Adj_H(S_{25}) = \{14, 20\}$. Figure 3 shows the improvement of the separator by making the move $Z = S_{13}$ (see the top two grids) and the move $Z = S_{25}$ (see the middle two grids).



An alternating breadth-first search tree is a special case of an *alternating breadth-first level structure*. Let X_0 be some initial set of exposed nodes in S; X_0 forms the first level. Define the next level $X_1 = Adj_H(X_0)$, namely, those vertices in $Border(B)$ adjacent to vertices in X_0. The next level X_2 contains all nodes in S that are mates

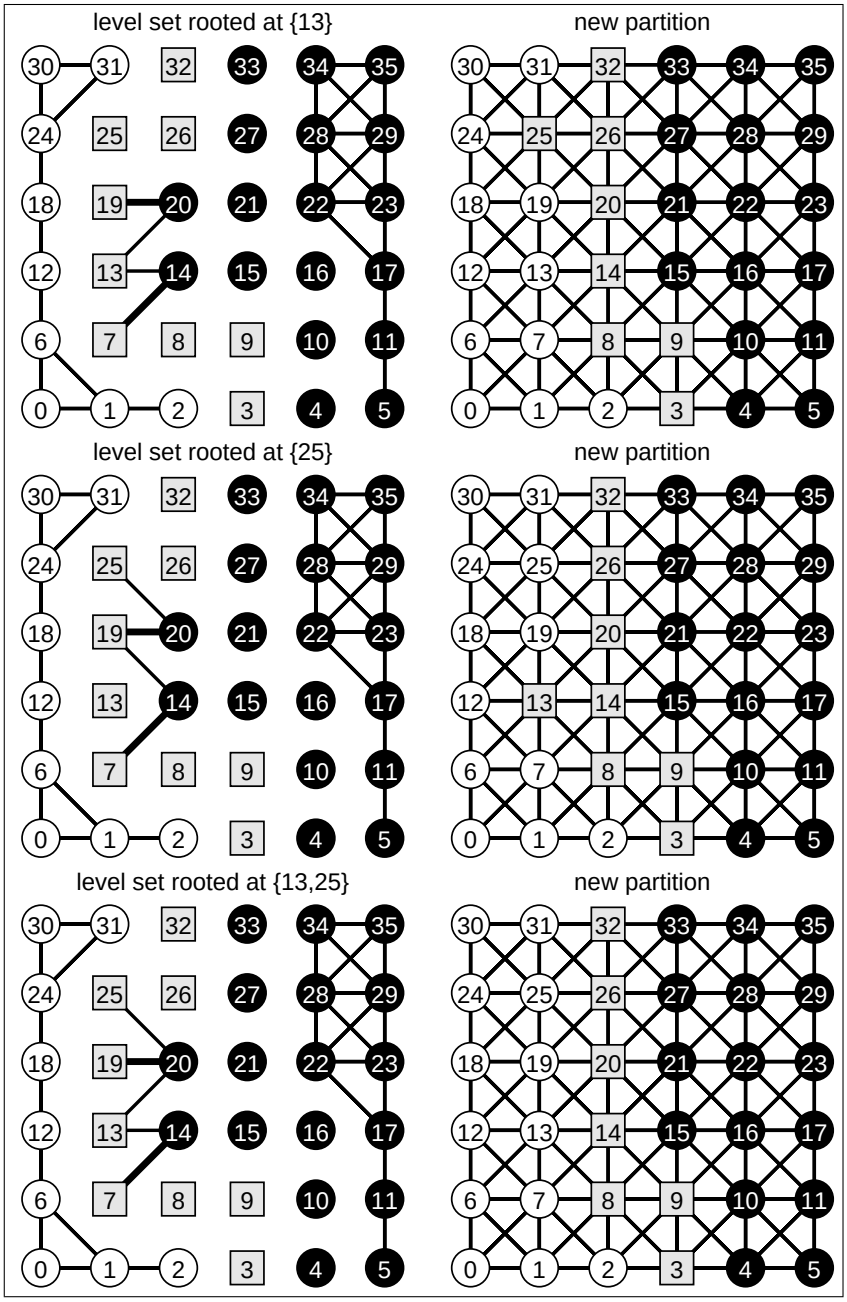


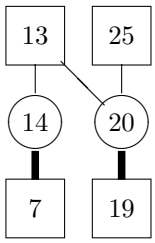
FIG. 3. *Alternating breadth-first level structures and their improved partitions.*

with nodes in X_1. In general, the level sets have the following form:

$$X_{2i} = \bigcup_{x \in X_{2i-1}} \text{mate}(x) \subseteq S,$$

$$X_{2i+1} = Adj_H \left(\bigcup_{j=0}^{2i} X_j \right) \subseteq Border(B).$$

The move set is $Z = X_0 \cup X_2 \cup \ldots$ while its boundary set is $Adj_H(Z) = X_1 \cup X_3 \cup \ldots$. For example, the alternating breadth-first level structure for $X_0 = \{13, 25\}$ is found below.



The move set is $S_{\{13,25\}} = \{7, 13, 19, 25\}$ while its boundary is $Adj_H(S_{\{13,25\}}) = \{14, 20\}$. Figure 3 shows the improvement of the separator by making the move $Z = S_{\{13,25\}}$ (see the bottom two grids). Note that the resulting separator is smaller than the separator induced by the two move sets S_{13} and S_{25}.

The first improvement to [24] is to use *all* exposed nodes in S to find a subset $Z \subseteq S$ that maximizes the decrease in separator size. It is based on the following extension [25], [26] of the result in Theorem 3.2 for separator-size reduction of greater than one.

THEOREM 3.3 (see Pothen and Fan [26]). *Define*

$$S_I = \{s \in S \mid s \text{ is reachable from some exposed node in } S \text{ via alternating paths}\}.$$

Then

- $|S_I| - |Adj_H(S_I)| > 0$,
- $|S_I| - |Adj_H(S_I)| = \max_{Z \subseteq S}\{|Z| - |Adj_H(Z)|\}$,
- S_I *is the smallest subset of* S *with this maximum value* $|S_I| - |Adj_H S_I|$.

The subset S_I can be constructed by performing an alternating breadth-first search starting with X_0, which contains *all* exposed nodes of S.

Theorems 3.2 and 3.3 provide the end points of a range of separator subsets with the size-improving property. Indeed, consider any subset X_0 of exposed nodes in S. It is easy to verify that the corresponding subset

$$Z = \bigcup \{S_x \mid x \in X_0\}$$

satisfies the condition $|Z| - |Adj_H(Z)| > 0$. This gives a number of choices in selecting a separator-improving subset. Although the subset S_I provides the maximum reduction in separator size, one might accept a smaller reduction in exchange for a better balance in the two components.

3.3. The Dulmage–Mendelsohn Decomposition. In [26], Pothen and Fan relate the subset S_I used in separator size reduction with the Dulmage–Mendelsohn decomposition of bipartite graphs [8]. The decomposition is also useful in our context in finding a balance-improving separator subset. Let $H(S, B)$ be the induced bipartite graph from a given partition $[S, B, W]$. Assume that a maximum matching M is given on H.

The *Dulmage–Mendelsohn decomposition* of S is the decomposition of S into three disjoint subsets: $S = S_I \cup S_R \cup S_X$, where

$$S_I = \{s \in S \mid s \text{ is reachable from some exposed node in } S \text{ via alternating paths}\},$$
$$S_X = \{s \in S \mid s \text{ is reachable from some exposed node in } B \text{ via alternating paths}\},$$
$$S_R = S \setminus (S_I \cup S_X).$$

Note we use the notation S_I to represent nodes reachable from internal exposed nodes, and S_X from external exposed nodes of S. S_R stands for the remaining nodes. We shall also use the notation $\langle S_I, S_X, S_R\rangle$ to represent the Dulmage–Mendelshohn decomposition of S. We now quote some results about this decomposition relevant to the partition improvement scheme.

THEOREM 3.4 (see Dulmage and Mendelsohn [8]). *The Dulmage–Mendelsohn decomposition $\langle S_I, S_X, S_R\rangle$ of S is independent of the maximum matching used to define the alternating paths.*

THEOREM 3.5 (see Pothen and Fan [26]). *The set $S_I \cup S_R$ satisfies the following:*

- $|S_I \cup S_R| - |Adj_H(S_I \cup S_R)| = |S_I| - |Adj_H(S_I)|$,
- *$S_I \cup S_R$ is the largest subset of S with the maximum value* $\max_{Z\subseteq S}\{|Z| - |Adj_H(Z)|\}$.

Theorem 3.3 states that S_I, if used, is the smallest subset of S with the maximum reduction $|S_I| - |Adj_H(S_I)|$ in separator size. On the other hand, Theorem 3.5 identifies $S_I \cup S_R$ as the largest subset with such maximum reduction in separator size. Moving S_I or $S_I \cup S_R$ will achieve the same amount of size reduction, but the balance for the resulting partition will be better for one or the other of the two moves.

By symmetry, there is a similar Dulmage–Mendelsohn decomposition $\langle B_I, B_X, B_R\rangle$ of B, the other part of the bipartite graph, where

$$B_I = \{b \in B \mid b \text{ is reachable from some exposed node in } B \text{ via alternating paths}\},$$
$$B_X = \{b \in B \mid b \text{ is reachable from some exposed node in } S \text{ via alternating paths}\},$$
$$B_R = B \setminus (B_I \cup B_X).$$

THEOREM 3.6 (see Dulmage and Mendelsohn [8]). *$S_X = Adj_H(B_I)$ and $B_X = Adj_H(S_I)$.*

The set S_X is given by the adjacent set of B_I, the set of reachable nodes in B from internal exposed nodes via alternating paths. The set B_I can be determined in the same way as S_I, by forming the alternating breadth-first search forest from the set of exposed nodes in B.

For the example in Figure 2, the sets of exposed nodes in S and B are $\{13, 25\}$ and $\{16, 33\}$, respectively. This gives the following:

$$\begin{aligned} S_I &= \{7, 13, 19, 25\}, & B_I &= \{4, 10, 16, 21, 27, 33\},\\ S_X &= \{3, 9, 26, 32\}, & B_X &= \{14, 20\},\\ S_R &= \{8\}, & B_R &= \{15\}. \end{aligned}$$

In Figure 4, we illustrate the Dulmage–Mendelsohn decomposition of the bipartite graph H of Figure 2. The six sets are arranged to illustrate their adjacency relationships.

It is instructive to interpret the decompositions $\langle S_I, S_X, S_R\rangle$ and $\langle B_I, B_X, B_R\rangle$ in connection with our partition improvement objective. For the given separator S, we can extend it to include its adjacent set in the B portion to obtain a *wide* separator $S \cup Border(B)$. The Dulmage–Mendelsohn decomposition provides machinery whereby a separator can be obtained from this wide separator, such that it is of minimum cover among all separator subsets of $S \cup Border(B)$. Indeed, it is clear that the following are two such separator subsets:

$$S_X \cup S_R \cup B_X, \qquad S_X \cup B_R \cup B_X.$$

Either one of them can be used to achieve a maximum reduction in separator size in the new partition.

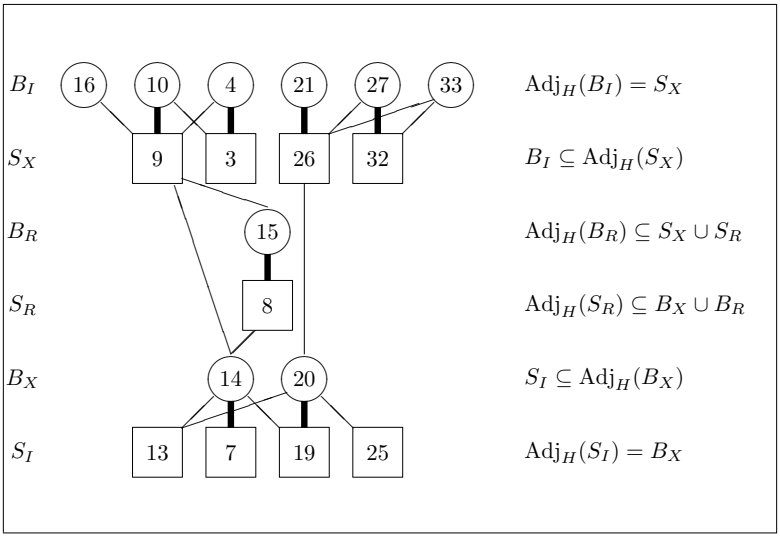


FIG. 4. *Dulmage–Mendelsohn decomposition.*

4. Using the Dulmage–Mendelsohn decomposition to improve balance.

4.1. Using the set S_R. In the discussion in the last section, we are looking for a separator-improving subset Z of S satisfying $|Z| > |Adj_H(Z) \cap B|$. If no such subset can be found, no reduction in separator size by graph matching is possible. In terms of the Dulmage–Mendelsohn decomposition, this is equivalent to the condition that the current separator S is already of minimum size among covering separator subsets of $S \cup Border(B)$. The algorithm as presented in [24] will terminate if there is no reduction in separator size via graph matching.

However, based on our evaluation function $\gamma(B, W, S)$, it may still be possible to improve the partition by reducing the imbalance ratio $\max\{|B|, |W|\}/\min\{|B|, |W|\}$. We can search for a subset Z of S with $|Adj_H(Z)| = |Z|$. A move of such a subset to the smaller portion W will replace Z by $Adj_H(Z)$ in S so that there will be no change in separator size. However, there may be a reduction in the imbalance.

When S_I is empty (implying that size reduction is not possible by this approach), the subset S_R can be used to reduce the imbalance. The next theorem contains an interesting property of this subset; to establish, we need the following lemma.

LEMMA 4.1. *Let $S_I = \emptyset$. Consider a subset Z of S. If $Z \cap S_X \neq \emptyset$, then $|Z| < |Adj_H(Z)|$.*

Proof. $S_I = \emptyset$ implies that there is a complete matching from S into B. By Theorem 3.1, $|Z| \leq |Adj_H(Z)|$ for every subset Z of S.

Let Z be a subset of S with $Z \cap S_X \neq \emptyset$. Assume for contradiction that $|Z| = |Adj_H(Z)|$. This means $Adj_H(Z)$ is exactly the set of matched vertices of Z for a given maximum matching. Let s be a vertex in $Z \cap S_X$. Then there exists an exposed vertex $b_e \in B$ and an alternating path from b_e to s

$$(b_e, s_1, b_1, \ldots, s_t, b_t, s_{t+1} = s),$$

where each pair $\{s_i, b_i\}$ belongs to the maximum matching. Let m be the smallest index such that $s_m \in Z$. If $m = 1$, this is a contradiction since $b_e \in Adj_H(s_1) \subseteq$

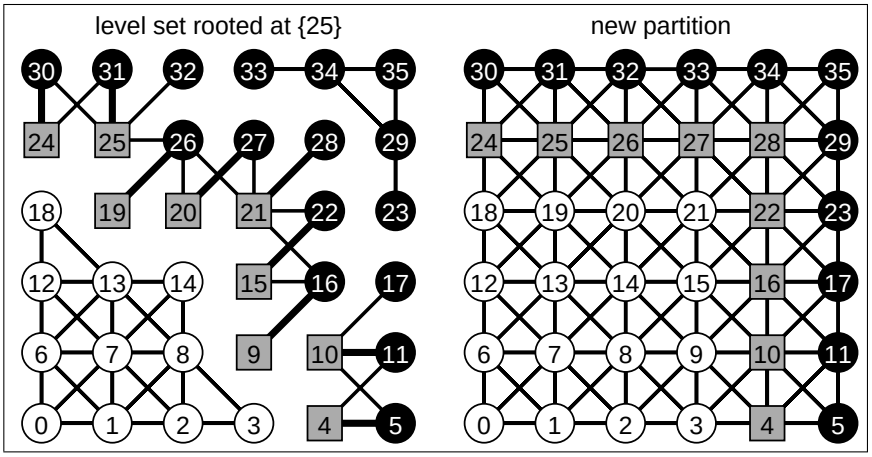


FIG. 5. *Improving the balance.*

$Adj_H(Z)$ and b_e does not have a mate in Z. For the case $m > 1$, this is again a contradiction since $b_{m-1} \in Adj_H(s_m) \subseteq Adj_H(Z)$ and the mate s_{m-1} of b_{m-1} is not in Z by the choice of m. Therefore, we have $|Z| < |Adj_H(Z)|$. □

THEOREM 4.2. *Let $S_I = \emptyset$. The separator subset S_R is the largest subset of S such that its size is the same as the size of its adjacent set.*

Proof. By Theorem 3.5 we have

$$|S_I \cup S_R| - |Adj_H(S_I \cup S_R)| = |S_I| - |Adj_H(S_I)|,$$

so that if $S_I = \emptyset$, $|S_R| - |Adj_H(S_R)| = 0$.

Consider any subset Z of S with the property $|Z| = |Adj_H(Z)|$. By Lemma 4.1, $Z \cap S_X$ is empty, which implies $Z \subseteq S_R$. □

Theorem 4.2 suggests that the subset S_R is the key to finding a balance-improving separator subset. We first note from [25], [26] that, in general, we have $|S_R| = |B_R|$. Furthermore, we have

$$Adj_H(S_I \cup S_R) = B_X \cup B_R,$$

so that when $S_I = \emptyset$, we have $B_X = \emptyset$ and $Adj_H(S_R) = B_R$. Therefore, when the separator subset S_I is empty, the move of S_R to W will give a new separator

$$S_{S_R \mapsto W} = (S \cup B_R) \setminus S_R,$$

so that $|S_{S_R \mapsto W}| = |(S \cup B_R) \setminus S_R| = |S|$.

Consider the example in Figure 5. There is a complete matching from the set S to B so that in the induced bipartite graph, $S_I = \emptyset$. This implies the separator-improving technique in the last section is not applicable. Note that the Dulmage–Mendelsohn decomposition is given by the following:

$$\begin{aligned} S_I &= \emptyset, & B_I &= \{5, 11, 17, 30, 31, 32\}, \\ S_X &= \{4, 10, 24, 25\}, & B_X &= \emptyset, \\ S_R &= \{9, 15, 19, 20, 21\}, & B_R &= \{16, 22, 26, 27, 28\}. \end{aligned}$$

For this example, moving the subset S_R from S to W will have the net effect of replacing it by B_R in S. In this way, the new separator will be $\{4, 10, 16, 22, 24, 25, 26, 27, 28\}$, which is the same size as before.

Now consider a separator subset Z with the property $|Z| = |Adj_H(Z)|$. Moving it to the portion W will preserve the separator size. The next result gives a simple necessary and sufficient condition for the move to improve the evaluation function $\gamma[S, B, W]$.

THEOREM 4.3. *Let $[S, B, W]$ be a given partition with $|B| \geq |W|$ and $S_I = \emptyset$. Consider a subset Z with $|Z| = |Adj_H(Z)|$. The move of the subset Z to W will reduce the evaluation function if and only if $|Z| < |B| - |W|$.*

Proof. Let

$$[S, B, W]_{Z \mapsto W} = [S_{Z \mapsto W}, B_{Z \mapsto W}, W_{Z \mapsto W}]$$

be the new partition after the move of the subset Z from S to W. It is clear that $|S_{Z \mapsto W}| = |S|$, $|B_{Z \mapsto W}| = |B| - |Z|$, and $|W_{Z \mapsto W}| = |W| + |Z|$.

Case 1. $|B_{Z \mapsto W}| \geq |W_{Z \mapsto W}|$.

$$\begin{aligned} \gamma[S, B, W] - \gamma[S, B, W]_{Z \mapsto W} &= |S| \left(1 + \alpha \frac{|B|}{|W|}\right) - |S| \left(1 + \alpha \frac{|B| - |Z|}{|W| + |Z|}\right) \\ &= \frac{\alpha |S|\ |Z|\ (|B| + |W|)}{|W|(|W| + |Z|)} > 0. \end{aligned}$$

Case 2. $|B_{Z \mapsto W}| < |W_{Z \mapsto W}|$.

$$\begin{aligned} \gamma[S, B, W] - \gamma[S, B, W]_{Z \mapsto W} &= |S| \left(1 + \alpha \frac{|B|}{|W|}\right) - |S| \left(1 + \alpha \frac{|W| + |Z|}{|B| - |Z|}\right) \\ &= \frac{\alpha |S|\ |Z|\ (|B| + |W|)(|B| - |W| - |Z|)}{|W|(|B| - |Z|)}. \end{aligned}$$

Assume $|Z| < |B| - |W|$. The evaluation function will be reduced in Case 1. Moreover, in Case 2, we have $|B| - |W| - |Z| > 0$ so $\gamma[S, B, W] - \gamma[S, B, W]_{Z \mapsto W} > 0$.

On the other hand, assume that $\gamma[S, B, W] - \gamma[S, B, W]_{Z \mapsto W} > 0$. In Case 1 we have $|B| - |Z| > |W| + |Z|$, which implies that

$$|B| - |W| > 2|Z| > |Z|.$$

Furthermore, in Case 2, a reduction in the evaluation function implies that $|B| - |W| - |Z| > 0$ or $|Z| < |B| - |W|$. ∎

By Theorems 4.2 and 4.3, to improve the balance of a given partition, we should be looking for a subset Z of S_R such that $|Z| = |Adj_H(Z)| < |B| - |W|$. Of course, if $|S_R| < |B| - |W|$, this set S_R is a good choice. Otherwise, we need to find proper subsets of S_R.

4.2. Finding balance-improving subsets of S_R. Finding a subset Z with $|Z| = |Adj_H(Z)|$ is related to the problem of reordering a sparse square matrix to block lower triangular form. In [26], Pothen and Fan provide an algorithm to compute the block triangular form of a sparse matrix. In their "fine decomposition" step, the square submatrix associated with the vertices in S_R and B_R are further reordered into block lower triangular form. (Pothen and Fan actually compute a block upper triangular form, but the algorithm can be adapted for block lower triangular form.) Their approach involves the following substeps:

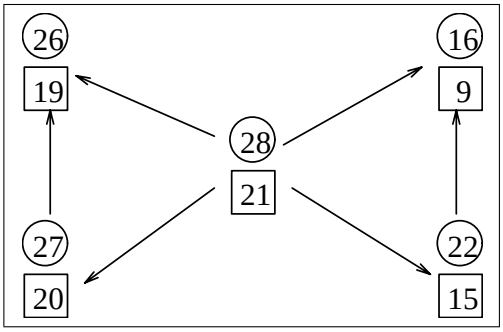


FIG. 6. *Induced directed graph.*

- Form a directed graph based on the bipartite subgraph of S_R and B_R. The directed graph consists of nodes from S_R. For two nodes x and y in S_R, there is a directed edge from x to y in this new directed graph if and only if there is an edge from x to the mate of y in B_R.
- Determine the strongly connected components of this directed graph. (The quotient graph using the strongly connected components forms a directed acyclic graph or, in short, a dag).
- Order the strongly connected components of this directed graph by a *reverse topological ordering* (i.e., an ordering of the nodes in the directed graph so that *all* the directed edges are pointing backwards to the left).

The reverse topological ordering of the strongly connected components of this directed graph will induce an ordering of the vertices in S_R and B_R so that the bipartite graph with this new reordering has a block lower triangular form. It should be clear from the block lower triangular structure that any subset Z of nodes of S_R corresponding to the leading blocks in the triangular form has this desirable property $|Z| = |Adj_H(Z)|$.

It is instructive to apply this scheme to the example of Figure 5. The new directed graph formed will consist of nodes from $S_R = \{9, 15, 19, 20, 21\}$. Figure 6 shows the directed graph; each vertex of this directed graph is labeled with both the node in S_R and its mate in B. There is no cycle in this directed graph, so that each node forms a strongly connected component. Furthermore, the following is a reverse topological ordering:

$$9, 15, 19, 20, 21$$

and the corresponding matrix is lower triangular:

$$\begin{array}{c} \\ 9 \\ 15 \\ 19 \\ 20 \\ 21 \end{array} \begin{array}{c} \begin{array}{ccccc} 16 & 22 & 26 & 27 & 28 \end{array} \\ \begin{pmatrix} \bullet & & & & \\ \bullet & \bullet & & & \\ & & \bullet & & \\ & & \bullet & \bullet & \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{pmatrix} \end{array}.$$

We can then deduce from this reverse topological ordering that all of the following subsets have the property $|Z| = |Adj_H(Z)|$:

$$\{9\}, \{9, 15\}, \{9, 15, 19\}, \{9, 15, 19, 20\}, \{9, 15, 19, 20, 21\}.$$

It is interesting to note that there are different reverse topological orderings of this directed graph. They will provide additional such subsets. For example, $\{19, 20, 9, 15, 21\}$ is a different reverse topological ordering, and the subsets $\{19\}, \{19, 20\}, \{19, 20, 9\}$ also have the size-preserving property.

5. Partition improvement on compressed graphs.

5.1. Compressed graphs. The Dulmage–Mendelsohn decomposition is the basic tool used in the last two sections to improve a given two-set partition. In this section, we explore efficient ways of computing this decomposition for some practical classes of matrix problems. It is common for graphs from applications to have sets of vertices with identical adjacency structures, e.g., in a finite element graph, a given geometric location may have multiple displacements and rotations. Such vertex pairs are sometimes referred to as *indistinguishable* in the sparse matrix research community. More formally, two vertices x and y are said to be *indistinguishable* if

$$Adj(x) \cup \{x\} = Adj(y) \cup \{y\}.$$

The notion of *compressed graph* is introduced in [1], [6], where each vertex of the compressed graph corresponds to (possibly) several indistinguishable vertices in the original graph. A compressed graph can be viewed as a *quotient graph* of the original unit-weight graph consisting of weighted compressed vertices. The motivation in [1] is for efficient implementations of some sparse matrix ordering algorithms. The number of vertices in a compressed graph can be many fewer than those of the original unit-weight graph. Since most graph algorithms have a strong $O(|V|)$ or $O(|E|)$ component to their complexity, it would be quite beneficial to work with a compressed graph instead of the original graph.

Following [1], we use boldface $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ to represent a compressed graph. A boldface $\mathbf{v}$ is used to denote a compressed vertex in $\mathbf{V}$ that corresponds to a set of indistinguishable vertices in V. For a given compressed graph, it is helpful to define its associated *compression* to be a mapping $\kappa : V \longrightarrow \mathbf{V}$, where $\kappa(v)$ is the compressed vertex in $\mathbf{V}$ containing the vertex v. This means $\kappa(v)$ is a subset of vertices in V (containing v) that are indistinguishable from v in G. However, we do not require $\kappa(v)$ to include *all* possible indistinguishable vertices of v. An edge $(\mathbf{u}, \mathbf{v})$ is in the compressed edge set $\mathbf{E}$ if $(\mathbf{u} \times \mathbf{v}) \cap E \neq \emptyset$. The theory and algorithm to be developed apply to any level of compression (partial or complete). Note that this is a *lossless* representation; that is, given $\mathbf{E}$ and κ, we can always recover the original edge set E.

We also extend the usage of κ to subsets: for a subset $Y \subseteq V$, $\kappa(Y) = \{\kappa(y) \mid y \in Y\} \subseteq \mathbf{V}$. For a compressed vertex $\mathbf{v} \in \mathbf{V}$, define its *weight* $wt(\mathbf{v})$ to be the number of indistinguishable vertices contained in $\mathbf{v}$. This notion can be extended to the weight of a subset of compressed vertices: for a subset $\mathbf{Y}$ of $\mathbf{V}$,

$$wt(\mathbf{Y}) = \sum \{wt(\mathbf{y}) \mid \mathbf{y} \in \mathbf{Y}\}.$$

We now consider the partition improvement techniques of the last two sections in the context of compressed graphs.

5.2. Definitions for bipartite compressed graphs. Let $\mathbf{G}$ be a given compressed graph with compression κ and a two-set partition $[\mathbf{S}, \mathbf{B}, \mathbf{W}]$; that is, $Adj(\mathbf{B}) \subseteq \mathbf{S}$ and $Adj(\mathbf{W}) \subseteq \mathbf{S}$. We first make a connection of this compressed partition with a partition on the original graph G.

THEOREM 5.1. *There is a unique two-set partition* $[\overline{S}, \overline{B}, \overline{W}]$ *on* G *such that* $\kappa(\overline{S}) = \mathbf{S}$, $\kappa(\overline{B}) = \mathbf{B}$, *and* $\kappa(\overline{W}) = \mathbf{W}$.

Proof. It is clear that the subsets defined by

$$\overline{S} = \{v \in V \mid \kappa(v) \in \mathbf{S}\},\ \overline{B} = \{v \in V \mid \kappa(v) \in \mathbf{B}\},\ \overline{W} = \{v \in V \mid \kappa(v) \in \mathbf{W}\}$$

satisfy the conditions $\kappa(\overline{S}) = \mathbf{S}$, $\kappa(\overline{B}) = \mathbf{B}$, and $\kappa(\overline{W}) = \mathbf{W}$, respectively. To prove that they form a two-set partition on G, it suffices to show that $Adj(\overline{B}) \subseteq \overline{S}$ (by symmetry, we have $Adj(\overline{W}) \subseteq \overline{S}$). This is the case since otherwise it would have implied that $\mathbf{W}$ and $\mathbf{B}$ are adjacent, contradicting the fact that $\mathbf{S}$ separates them. The uniqueness follows from the fact that $[\mathbf{S}, \mathbf{B}, \mathbf{W}]$ is a partition on $\mathbf{G}$. ∎

The simple connection in Theorem 5.1 allows the partition improvement techniques described in sections 3 and 4 using the Dulmage–Mendelsohn decomposition to be applied to the induced partition $[\overline{S}, \overline{B}, \overline{W}]$ of the unit-weight graph. We now show that the decomposition of the original graph can be readily obtained from a similar decomposition of the compressed graph. We first extend the various notions used in the formulation of the Dulmage–Mendelsohn decomposition from unit-weight graphs to compressed graphs.

5.2.1. Compressed matching and maximum matching. As before, let $[\mathbf{S}, \mathbf{B}, \mathbf{W}]$ be a given partition on the compressed graph $\mathbf{G}$. This will in turn define a bipartite compressed graph $\mathbf{H}(\mathbf{S}, \mathit{Border}(\mathbf{B}))$. We extend the notion of matching to bipartite compressed graphs. In the unit-weight case, a matching can be viewed as an assignment of an integer value $f(s, b)$ to each edge (s, b) in the bipartite graph such that

- for every edge (s, b), $f(s, b) \geq 0$;
- for every vertex $\tilde{s} \in S$, $1 \geq \sum\{f(\tilde{s}, b) \mid b \in B\}$;
- for every vertex $\tilde{b} \in B$, $1 \geq \sum\{f(s, \tilde{b}) \mid s \in S\}$.

It follows that the assigned values can either be zero (unmatched) or one (matched). Furthermore, a maximum matching is one that will maximize the sum $\sum\{f(s, b) \mid s \in S,\ b \in B\}$, which is the number of edges in the matching.

With this interpretation, we generalize a compressed matching of a bipartite compressed graph to be an assignment of integer values $f(\mathbf{s}, \mathbf{b})$ to the edges such that they satisfy the following three conditions:

- for every edge $(\mathbf{s}, \mathbf{b})$, $f(\mathbf{s}, \mathbf{b}) \geq 0$;
- for every compressed vertex $\tilde{\mathbf{s}} \in \mathbf{S}$, $wt(\tilde{\mathbf{s}}) \geq \sum\{f(\tilde{\mathbf{s}}, \mathbf{b}) \mid \mathbf{b} \in \mathbf{B}\}$;
- for every compressed vertex $\tilde{\mathbf{b}} \in \mathbf{B}$, $wt(\tilde{\mathbf{b}}) \geq \sum\{f(\mathbf{s}, \tilde{\mathbf{b}}) \mid \mathbf{s} \in \mathbf{S}\}$.

Note that we have used the weight of each vertex instead of unit weight in the sums above. A maximum compressed matching is one that maximizes the total edge value:

$$\sum\{f(\mathbf{s}, \mathbf{b}) \mid \mathbf{s} \in \mathbf{S},\ \mathbf{b} \in \mathbf{B}\}.$$

Compressed exposed nodes and alternating paths. In the unit-weight bipartite graph, an exposed node s_e is one such that none of its incident edges belong to the matching. In terms of the value $f(s, b)$, this is equivalent to the condition that for the node s_e

$$1 > \sum\{f(s_e, b) \mid b \in B\} = 0.$$

In a compressed bipartite graph, we define an exposed node $\mathbf{s}_e \in \mathbf{S}$ to be a node such that

$$wt(\mathbf{s}_e) > \sum\{f(\mathbf{s}_e, \mathbf{b}) \mid \mathbf{b} \in \mathbf{B}\}.$$

The *exposure* of $\mathbf{s}$ is defined to be $wt(\mathbf{s}) - \sum\{f(\mathbf{s},\mathbf{b}) \mid \mathbf{b} \in \mathbf{B}\}$. Therefore, the exposure of an exposed node is positive. Exposed nodes and exposure are similarly defined for compressed vertices in $\mathbf{B}$.

For a given compressed matching $\mathbf{M}$ of the compressed bipartite graph, consider a path

$$\mathbf{s}_0 \longrightarrow \mathbf{b}_1 \longrightarrow \mathbf{s}_1 \longrightarrow \mathbf{b}_2 \longrightarrow \mathbf{s}_2 \longrightarrow \ldots \longrightarrow \mathbf{b}_m \longrightarrow \mathbf{s}_m \longrightarrow \ldots .$$

It is a compressed alternating path with respect to $\mathbf{M}$ if the alternate edges

$$(\mathbf{b}_1,\mathbf{s}_1), (\mathbf{b}_2,\mathbf{s}_2), \ldots, (\mathbf{b}_m,\mathbf{s}_m), \ldots$$

all have positive edge values from the matching $\mathbf{M}$. In such a case, we use the following to represent an alternating path:

$$\mathbf{s}_0 \longrightarrow \mathbf{b}_1 \Longrightarrow \mathbf{s}_1 \longrightarrow \mathbf{b}_2 \Longrightarrow \mathbf{s}_2 \longrightarrow \ldots \longrightarrow \mathbf{b}_m \Longrightarrow \mathbf{s}_m \longrightarrow \ldots ,$$

where a double-lined arrow is used to indicated an edge with positive edge value. An alternating path that starts with a compressed node from $\mathbf{B}$ is similarly defined.

5.3. Decomposition in bipartite compressed graphs. Recall that our objective is to improve a partition using the Dulmage–Mendelsohn decomposition, and we want to take advantage of compression in finding such decomposition. As before, let $[\mathbf{S},\mathbf{B},\mathbf{W}]$ be a given partition on the compressed graph $\mathbf{G}$ and $\mathbf{H}(\mathbf{S}, \mathit{Border}(\mathbf{B}))$ be the corresponding bipartite compressed graph. Furthermore, let $\mathbf{M}$ be a maximum compressed matching on $\mathbf{H}$. Consider the decomposition $\mathbf{S}_I \cup \mathbf{S}_R \cup \mathbf{S}_X$ of $\mathbf{S}$, where

$$\begin{aligned}
\mathbf{S}_I &= \{\mathbf{s} \in \mathbf{S} \mid \mathbf{s} \text{ is reachable from some } \textbf{exposed} \text{ node in } \mathbf{S} \text{ via } \textbf{alternating} \text{ paths}\},\\
\mathbf{S}_X &= \{\mathbf{s} \in \mathbf{S} \mid \mathbf{s} \text{ is reachable from some } \textbf{exposed} \text{ node in } \mathbf{B} \text{ via } \textbf{alternating} \text{ paths}\},\\
\mathbf{S}_R &= \mathbf{S} \setminus (\mathbf{S}_I \cup \mathbf{S}_X).
\end{aligned}$$

Note that we have used the boldface **exposed** and **alternating** in the decomposition above to emphasize the use of the extended definitions of compressed exposed nodes and compressed alternating paths for compressed bipartite graphs. The decomposition $\mathbf{B}_I \cup \mathbf{B}_R \cup \mathbf{B}_X$ of $\mathbf{B}$ can be similarly defined.

We now make the connection of this decomposition with the Dulmage–Mendelsohn decomposition in the unit-weight graph. Consider the unique partition $[\overline{S},\overline{B},\overline{W}]$ of the unit-weight graph G satisfying $\kappa(\overline{S}) = \mathbf{S}$, $\kappa(\overline{B}) = \mathbf{B}$, and $\kappa(\overline{W}) = \mathbf{W}$ in Theorem 5.1. This partition will in turn determine a unit-weight bipartite graph $\overline{H}(\overline{S},\overline{B})$. We now relate the decomposition $\mathbf{S}_I \cup \mathbf{S}_R \cup \mathbf{S}_X$ with the Dulmage–Mendelsohn decomposition of $\overline{S}$ in this $\overline{H}(\overline{S},\overline{B})$. Note that the Dulmage–Mendelsohn decomposition $\langle \overline{S}_I, \overline{S}_X, \overline{S}_R \rangle$ of $\overline{S}$ is independent of any maximum matching used in $\overline{H}$. But in order to make the connection between the two decompositions, we need to define an induced matching of $\overline{H}$ from $\mathbf{H}$.

Let $\mathbf{M}$ be a compressed matching on $\mathbf{H}(\mathbf{S}, \mathit{Border}(\mathbf{B}))$. An induced matching $\overline{M}$ on the unit-weight bipartite graph $\overline{H}(\overline{S},\overline{B})$ can be defined as follows. For each compressed vertex $\mathbf{s}$, we have the size condition

$$wt(\mathbf{s}) \geq \sum\{f(\mathbf{s},\mathbf{b}) \mid \mathbf{b} \in \mathbf{B}\}.$$

Therefore, for each incident edge $(\mathbf{s},\mathbf{b})$ we can always assign $f(\mathbf{s},\mathbf{b})$ distinct $\overline{S}$-vertices from $\mathbf{s}$ to this compressed edge. Similar allotments of $\overline{B}$-vertices from compressed vertices of $\mathbf{B}$ to compressed incident edges can be assigned.

Now for each compressed edge $(\mathbf{s}, \mathbf{b})$ with value $f(\mathbf{s}, \mathbf{b})$ in the matching $\mathbf{M}$, there will be $f(\mathbf{s}, \mathbf{b})$ $\overline{S}$-vertices from $\mathbf{s}$ and the same number of $\overline{B}$-vertices from $\mathbf{b}$ assigned to this edge. Since each $\overline{S}$-vertex in $\mathbf{s}$ is adjacent to each $\overline{B}$-vertex in $\mathbf{b}$, we can get $f(\mathbf{s}, \mathbf{b})$ different edges consisting of pairs of adjacent assigned vertices from $\mathbf{s}$ and $\mathbf{b}$. We place them in the set $\overline{M}$.

After doing this for every compressed edge, we see that the set of edges in $\overline{M}$, by construction, does not have common vertices. This means that the set $\overline{M}$ forms a matching. Furthermore, this set $\overline{M}$ satisfies the following property.

LEMMA 5.2. *Given* $\mathbf{M}$ *is a maximum matching on the compressed bipartite graph* $\mathbf{H}$, $\overline{M}$ *is a maximum matching on* $\overline{H}$.

Proof. Assume for contradiction that $\overline{M}$ is not maximum. We can therefore find an alternating path connecting two exposed vertices, say $\overline{s} \in \overline{S}$ and $\overline{b} \in \overline{B}$ (such a path is usually referred to as an augmenting path):

$$\overline{s} \longrightarrow \overline{b}_1 \Longrightarrow \overline{s}_1 \longrightarrow \ldots \longrightarrow \overline{b}_t \Longrightarrow \overline{s}_t \longrightarrow \overline{b}.$$

Through compression, this corresponds to a path or walk in the compressed graph

$$\kappa(\overline{s}) \longrightarrow \kappa(\overline{b}_1) \Longrightarrow \kappa(\overline{s}_1) \longrightarrow \ldots \longrightarrow \kappa(\overline{b}_t) \Longrightarrow \kappa(\overline{s}_t) \longrightarrow \kappa(\overline{b}).$$

Since $\overline{s} \in \kappa(\overline{s})$ and $\overline{b} \in \kappa(\overline{b})$ are both exposed in $\overline{M}$, we can increase the total matching in $\mathbf{M}$ by at least one by alternately increasing and decreasing the f values along this path. This contradicts the fact that $\mathbf{M}$ is a maximum matching on $\mathbf{H}$. ∎

By this lemma, $\overline{M}$ is a maximum matching on $\overline{H}$ so that the Dulmage–Mendelsohn decomposition $\langle \overline{S}_I, \overline{S}_X, \overline{S}_R \rangle$ of $\overline{S}$ can be determined using $\overline{M}$.

THEOREM 5.3. $\kappa(\overline{S}_I) = \mathbf{S}_I$, $\kappa(\overline{S}_R) = \mathbf{S}_R$, $\kappa(\overline{S}_X) = \mathbf{S}_X$.

Proof. We only prove $\kappa(\overline{S}_I) = \mathbf{S}_I$ and leave the remaining two for the readers. We first show $\kappa(\overline{S}_I) \subseteq \mathbf{S}_I$. Consider a compressed vertex $\kappa(\overline{s}) \in \kappa(\overline{S}_I)$ with $\overline{s} \in \overline{S}_I$. This means that there exists an alternating path

$$s_0 \longrightarrow b_1 \Longrightarrow s_1 \longrightarrow \ldots \longrightarrow b_t \Longrightarrow s_t = \overline{s}$$

from some exposed node $s_0 \in \overline{S}$. This induces a compressed alternating path or walk in $\mathbf{H}$:

$$\kappa(s_0) \longrightarrow \kappa(b_1) \Longrightarrow \kappa(s_1) \longrightarrow \ldots \longrightarrow \kappa(b_t) \Longrightarrow \kappa(s_t) = \kappa(\overline{s})$$

and $\kappa(s_0)$ is exposed in $\mathbf{S}$. Therefore, $\kappa(\overline{s}) \in \mathbf{S}_I$.

We now show $\mathbf{S}_I \subseteq \kappa(\overline{S}_I)$. Consider a compressed node $\mathbf{s} \in \mathbf{S}_I$. There exists an alternating path in $\mathbf{H}$

$$\mathbf{s}_0 \longrightarrow \mathbf{b}_1 \Longrightarrow \mathbf{s}_1 \longrightarrow \ldots \longrightarrow \mathbf{b}_t \Longrightarrow \mathbf{s}_t = \mathbf{s}$$

from an exposed $\mathbf{s}_0$. Choose a $\overline{s}_0 \in \mathbf{s}_0$, that is exposed in $\overline{M}$. For $0 < i < t$, choose a matched edge $(\overline{b}_i, \overline{s}_i)$ in $\overline{M}$, where $\kappa(\overline{b}_i) = \mathbf{b}_i$ and $\kappa(\overline{s}_i) = \mathbf{s}_i$; one is guaranteed since $f(\mathbf{b}_i, \mathbf{s}_i) > 0$. This forms an alternating path in $\overline{M}$ from an exposed node $\overline{s}_0$ to $\overline{s}_t$; therefore, $\overline{s}_t \in \overline{S}_I$. The result follows since $\kappa(\overline{s}_t) = \mathbf{s}$. ∎

The next two theorems follow directly from Theorem 5.3. The first theorem states that the weight of $\mathbf{S}_I$ is less than that of $\mathbf{B}_X$ so that the partition $[\mathbf{S}, \mathbf{B}, \mathbf{W}]$ can be improved by replacing the subset $\mathbf{S}_I$ with its adjacent set $\mathbf{B}_X$. The second theorem relates the weights of the two subsets $\mathbf{S}_R$ and $\mathbf{B}_R$.

THEOREM 5.4. *If* $\mathbf{S}_I$ *is nonempty, then* $wt(\mathbf{S}_I) < wt(\mathbf{B}_X)$.

THEOREM 5.5. $wt(\mathbf{S}_R) = wt(\mathbf{B}_R)$.

6. Partition improvement by maximum network flow.

6.1. Bipartite compressed graph matching by maximum network flow.

Finding an assignment of edge values to a bipartite compressed graph that corresponds to a maximum matching can be reformulated into a much-studied combinatorial problem involving flow through a network [9], [11], [12], [18], [28]. A *network* is a weighted directed graph with two special nodes: one with no incoming edges (the *source*), and one with no outgoing edges (the *sink*). There are capacity constraints associated with the edges and vertices. Most discussions of the network flow problem in the literature assume the use of edge capacities. The generalization to include both edge and vertex capacity is well known (for example, [22, pp. 120–121]). For our purposes, we only need to consider networks with finite vertex capacities, i.e., each vertex y is given a nonnegative integer value $\mathit{capacity}(y)$, called the *capacity* of the vertex. All edges have infinite capacity.

A *flow* is a function that assigns a nonnegative integer value $\mathit{flow}(y, z)$ to each directed edge (y, z). The flow satisfies two conditions:

- the amount of *in-flow* equals the amount of *out-flow* at each vertex except the source and sink;
- the in-flow must be within capacity of each vertex.

Let $\mathit{inflow}(y)$ denote the amount of in-flow into the vertex y, that is,

$$\mathit{inflow}(y) = \sum \{\mathit{flow}(v, y) \mid v \in V\}$$

and let $\mathit{outflow}(y)$ be the amount of out-flow from the vertex y,

$$\mathit{outflow}(y) = \sum \{\mathit{flow}(y, v) \mid v \in V\}.$$

For every vertex aside from the source and sink, the flow function satisfies

$$\mathit{inflow}(y) = \mathit{outflow}(y) \leq \mathit{capacity}(y).$$

We shall also refer to $\mathit{inflow}(y)$ (or equivalently $\mathit{outflow}(y)$) as the flow across the vertex y. A vertex y is said to be *saturated* or at capacity if $\mathit{inflow}(y) = \mathit{capacity}(y)$; otherwise, it is said to be below capacity or to have excess capacity. By convention, the source and sink have infinite capacity.

The value of the flow is the amount of out-flow from the source node, called $\mathit{outflow}(\mathbf{source})$[2], or equivalently, the amount of in-flow into the sink node, $\mathit{inflow}(\mathbf{sink})$. The *network flow problem* is to find a flow with the maximum value for a given network. It should be emphasized that we consider only integer capacity and flow values.

We now describe a network flow problem that when solved will give a solution to the maximum matching problem for bipartite compressed graphs. As before, let $\mathbf{H}(\mathbf{S}, \mathit{Border}(\mathbf{B}))$ be our bipartite compressed graph with weight function $\mathit{wt}(*)$. A bipartite network is constructed as follows:

- In addition to the **source** and **sink**, the nodes in the network are the vertices in $\mathbf{S}$ and $\mathit{Border}(\mathbf{B})$.
- For each vertex $\mathbf{s} \in \mathbf{S}$, add the directed edge $(\mathbf{source}, \mathbf{s})$ to the network.
- For each vertex $\mathbf{b} \in \mathit{Border}(\mathbf{B})$, add the directed edge $(\mathbf{b}, \mathbf{sink})$ to the network.

[2]We use boldface for **source** and **sink** to emphasize that we are working on the weighted compressed graph.

- For each edge $(\mathbf{s}, \mathbf{b})$, $\mathbf{s} \in \mathbf{S}$, $\mathbf{b} \in \mathit{Border}(\mathbf{B})$, in the graph $\mathbf{H}(\mathbf{S}, \mathit{Border}(\mathbf{B}))$, add a directed edge $(\mathbf{s}, \mathbf{b})$ to the network, where flow is assumed to go from $\mathbf{s}$ to $\mathbf{b}$ along this edge.
- All edges have infinite capacity. For each vertex $\mathbf{y}$ in $\mathbf{H}(\mathbf{S}, \mathit{Border}(\mathbf{B}))$, we set $\mathit{capacity}(\mathbf{y}) = \mathit{wt}(\mathbf{y})$.

In the top network of Figure 7 we illustrate the bipartite network obtained for the separator example of Figure 2. Arrows are used on edges to indicate the direction of flow (except for those involving the **source** and the **sink**). Edges with positive (zero) flow are thick (thin) lines. Note that there is a directed edge from the source to every vertex in $\mathbf{S}$ (the set of "square" vertices) and one from every vertex in $\mathit{Border}(\mathbf{B})$ (the set of "circle" vertices) to the sink.

We shall use the notation $\mathcal{N}_b$ (b for bipartite) to represent this bipartite network. To establish the equivalence between a max-flow solution on this bipartite network with a maximum matching on the bipartite compressed graph, we use the equivalence of *flow augmenting paths* in the former with *augmenting paths* in the latter. An augmenting path in the bipartite compressed graph is an alternating path whose first and last vertices are exposed in $\mathbf{S}$.

It is simple to generalize such augmenting paths for bipartite network flows. Indeed, a flow augmenting path for a bipartite network is a sequence of edges from the source to the sink with alternate *forward* and *backward* edges:

$$\textbf{source} \longrightarrow \mathbf{v}_1 \longrightarrow \mathbf{v}_2 \longleftarrow \mathbf{v}_3 \longrightarrow \mathbf{v}_4 \longleftarrow \ldots \longleftarrow \mathbf{v}_k \longrightarrow \textbf{sink}.$$

Furthermore, each backward edge $(\mathbf{v}_{2j+1}, \mathbf{v}_{2j})$ has positive flow and the vertices $\mathbf{v}_1$ and $\mathbf{v}_k$ are below capacity. It is easy to relate this with an augmenting path in the original bipartite compressed graph. Since $\mathbf{v}_1$ and $\mathbf{v}_k$ are below capacity, they are exposed in the graph matching. Any backward edge with positive flow means the two incident vertices are matched.

Since we will be considering flows on a general network, we must further generalize the notion of a flow augmenting path. When edges have finite capacity, a flow augmenting path is a path from the source to the sink such that forward edges are below capacity and backward edges have positive flow. In our networks the edges have infinite capacity and the vertices have finite capacity, so a flow augmenting path is a sequence of vertices $\langle \textbf{source} = \mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{v}_{k+1} = \textbf{sink} \rangle$ with these four properties.

- Two consecutive vertices $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ are connected by an edge in the network; a forward edge is of the form $(\mathbf{v}_i, \mathbf{v}_{i+1})$, a backward edge is of the form $(\mathbf{v}_{i+1}, \mathbf{v}_i)$.
- Any two consecutive forward edges $(\mathbf{v}_{i-1}, \mathbf{v}_i)$ and $(\mathbf{v}_i, \mathbf{v}_{i+1})$ implies vertex $\mathbf{v}_i$ is below capacity.
- Any backward edge $(\mathbf{v}_{i+1}, \mathbf{v}_i)$ has nonzero flow, i.e., $\mathit{flow}(\mathbf{v}_{i+1}, \mathbf{v}_i) > 0$.
- A vertex may appear in the path once or twice, via a forward edge, a backward edge, or both[3].

The overall flow value can be increased by increasing flow along the forward edges and decreasing flow along the backward edges.

[3]Technically speaking, if a vertex is visited twice we have a flow augmenting *walk*. Had we taken the more conventional route of handling vertex capacities by expanding a vertex $\mathbf{v}$ into a pair of vertices connected by an edge $(\mathbf{v}^-, \mathbf{v}^+)$ whose capacity is the weight of the vertex, then $\mathbf{v}^-$ would be visited by a forward edge, $\mathbf{v}^+$ would be visited by a forward or a backward edge, and there would be no repeated vertices along the path.

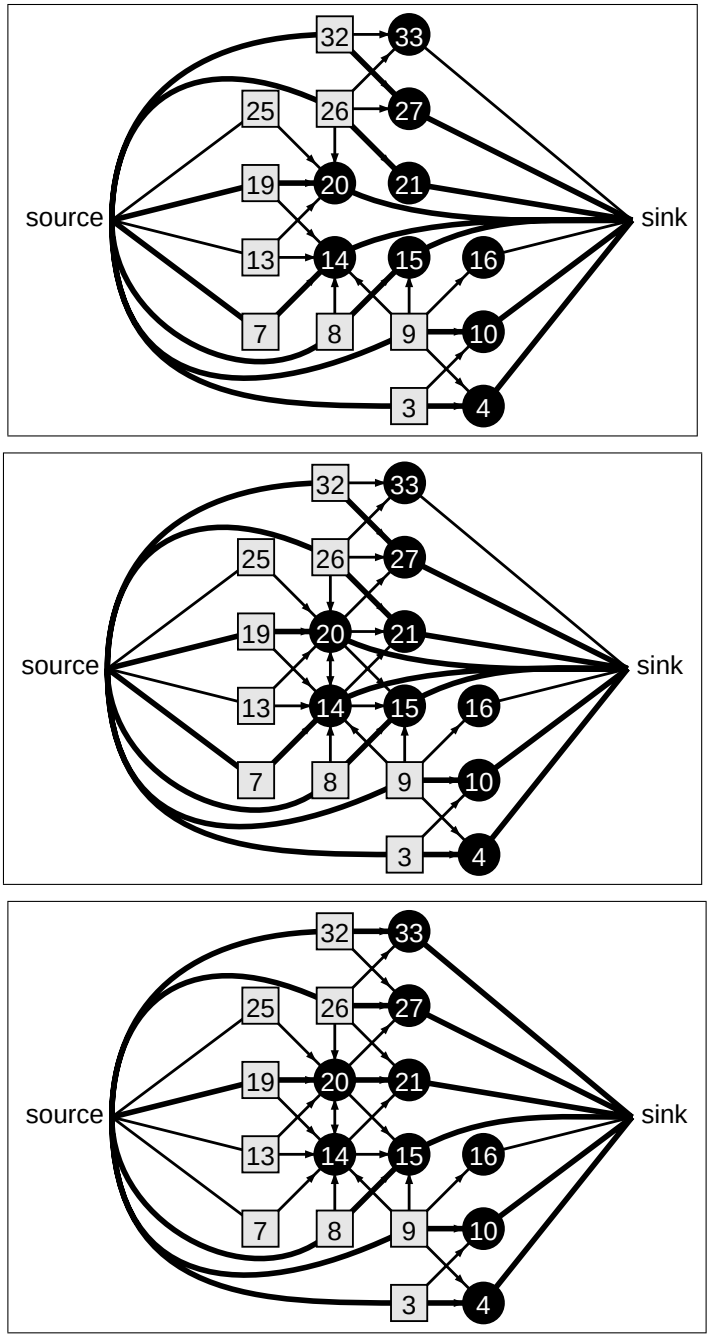


FIG. 7. *Top: $\mathcal{N}_b$, the original bipartite network used to find the Dulmage–Mendelsohn decomposition; middle: $\mathcal{N}_m$, the intermediate network found by adding edges that do not increase the max-flow; bottom: $\mathcal{N}_w$, the final three-layer network found by deleting edges from the middle layer vertices to the sink.*

6.2. Min-cut in network flow. The dual to the network max-flow is a *min-cut*. In our context of networks with finite vertex capacities and infinite edge capacities, a *cut* is a set of vertices whose removal separates the **source** from the **sink**, i.e., a separator of the graph from which the network was derived. A *min-cut* is a cut such

that its size

$$\sum\{capacity(v) \mid v \text{ belongs to the cut}\}$$

is minimum among all cuts. The well-known max-flow min-cut theorem states that the size of a min-cut is the same as the value of a max-flow.

It is interesting to relate min-cuts with the Dulmage–Mendelsohn decomposition. For a bipartite compressed graph, once we find a maximum matching we can determine the Dulmage–Mendelsohn decomposition and thus construct one or more minimum cover separators, such as $\mathbf{S}_X \cup \mathbf{S}_R \cup \mathbf{B}_X$ and $\mathbf{S}_X \cup \mathbf{B}_R \cup \mathbf{B}_X$. A covering separator of minimum size is equivalent to a *min-cut* of a bipartite network constructed from $\mathbf{S}$ and $Border(\mathbf{B})$.

There are two specific min-cuts of the network that are of interest. The tool we use is a *flow alternating path*. A flow alternating path differs from a flow augmenting path in that it need not start from the **source** nor end at the **sink**. Therefore, any contiguous sequence of edges from a flow augmenting path is a flow alternating path. We can now define the following subset:

$$\mathbf{R}_{\mathbf{source}} = \{\mathbf{v} \in \mathbf{V} \mid \mathbf{v} \text{ is reachable from } \mathbf{source} \text{ via a flow alternating path}\}.$$

Intuitively, the subset $\mathbf{R}_{\mathbf{source}}$ provides the "bottleneck" that limits the total flow to its present value. Indeed, the border of $\mathbf{R}_{\mathbf{source}}$ is a min-cut of the network. A similar subset can be defined with respect to the sink:

$$\mathbf{R}_{\mathbf{sink}} = \{\mathbf{v} \in \mathbf{V} \mid \text{the } \mathbf{sink} \text{ is reachable from } \mathbf{v} \text{ via a flow alternating path}\}.$$

The border of $\mathbf{R}_{\mathbf{sink}}$ is a min-cut of the network. For the network at the top of Figure 7, the two reach sets and their borders are given below.

$$\begin{aligned}
\mathbf{R}_{\mathbf{source}} &= \{3, 7, 8, 9, 13, 14, 19, 20, 25, 26, 32\},\\
Border(\mathbf{R}_{\mathbf{source}}) &= \{3, 8, 9, 14, 20, 26, 32\},\\
\mathbf{R}_{\mathbf{sink}} &= \{3, 4, 9, 10, 14, 15, 16, 20, 21, 26, 27, 32, 33\},\\
Border(\mathbf{R}_{\mathbf{sink}}) &= \{3, 9, 14, 15, 20, 26, 32\}.
\end{aligned}$$

In the context of the Dulmage–Mendelsohn decomposition, $\mathbf{R}_{\mathbf{source}} = \mathbf{S}_I \cup \mathbf{B}_X \cup \mathbf{S}_R \cup \mathbf{S}_X$, $Border(\mathbf{R}_{\mathbf{source}}) = \mathbf{B}_X \cup \mathbf{S}_R \cup \mathbf{S}_X$, $\mathbf{R}_{\mathbf{sink}} = \mathbf{B}_I \cup \mathbf{S}_X \cup \mathbf{B}_R \cup \mathbf{B}_X$, and $Border(\mathbf{R}_{\mathbf{sink}}) = \mathbf{S}_X \cup \mathbf{B}_R \cup \mathbf{B}_X$.

6.3. Enhancement techniques by network flow. In this section, we consider new partition improvement techniques based on network flows. We first consider a motivating example. Consider again the grid at the bottom of Figure 3. Using the Dulmage–Mendelsohn decomposition, we can determine the move set $\mathbf{S}_I = \{7, 13, 19, 25\}$ that decreases the separator size the most. The size of $S_X \cup B_X \cup B_X$, the new separator $\{3, 9, 14, 15, 20, 26, 32\}$, is seven. On the other hand, consider the two grids in Figure 8. The left-hand grid shows a wide separator $\mathbf{S} \cup Border(\mathbf{B})$ that contains 18 nodes. The right-hand grid shows a separator subset of size six, smaller than the "best" separator that was found using the Dulmage–Mendelsohn decomposition.

There is no contradiction here, yet there is a subtle point that needs to be understood. Theorem 3.3 states that $\mathbf{S}_I$ is the smallest subset of $\mathbf{S}$ that if absorbed by $\mathbf{W}$ will result in the largest decrease of separator size. The "move" that generated the partition in the right-hand grid of Figure 8 had $\mathbf{W}$ absorb the separator vertices

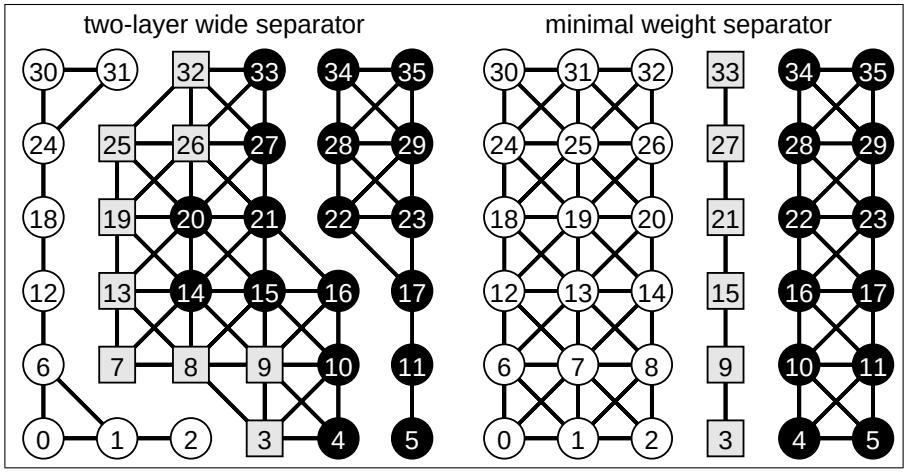


FIG. 8. *A two-layer wide separator and its minimal weight separator subset.*

$\{7, 8, 13, 19, 25, 26, 32\}$, but $\mathbf{W}$ also absorbed the black vertices $\{14, 20\}$, so it is a more general move than that covered by Theorem 3.3. Indeed, $\{7, 8, 13, 14, 19, 20, 25, 26, 32\}$ is the smallest subset of $\mathbf{S} \cup \mathit{Border}(\mathbf{B})$, which when moved to $\mathbf{W}$ will result in the largest decrease in separator size.

We first offer an intuitive explanation to the enhancement. Our goal is to improve an initial partition $[\mathbf{S}, \mathbf{B}, \mathbf{W}]$ of a given compressed graph. The separator $\mathbf{S}$ is first used to construct a compressed bipartite graph based on $\mathbf{S}$ and its adjacent set $\mathit{Border}(\mathbf{B})$ in $\mathbf{B}$. In section 6.1, we construct a bipartite network $\mathcal{N}_b$ based on this compressed bipartite graph. A max-flow min-cut solution to this bipartite network $\mathcal{N}_b$ can then be used to obtain an improved new partition for the original compressed graph.

We shall modify our bipartite network so that the max-flow value (and hence min-cut size) of the new network is no larger and possibly smaller. More importantly, the min-cut of this new network also corresponds to a separator of the underlying compressed graph. There is potential to obtain a smaller separator than the one from the original bipartite network.

We now describe how to construct the new network. Let $\mathbf{S} \cup \mathit{Border}(\mathbf{B})$ be the wide separator induced from $\mathbf{S}$. We have a new partition $[\mathbf{S} \cup \mathit{Border}(\mathbf{B}), \mathit{Int}(\mathbf{B}), \mathbf{W}]$. The wide separator has two portions $\mathbf{S}$ and $\mathit{Border}(\mathbf{B})$. Consider a further subdivision of the subset $\mathit{Border}(\mathbf{B})$ into

$$\mathbf{Y} = \{\mathbf{b} \in \mathit{Border}(\mathbf{B}) \mid \mathit{Adj}(\mathbf{b}) \cap \mathit{Int}(\mathbf{B}) = \emptyset\} \quad \text{and} \quad \mathbf{Z} = \mathit{Border}(\mathbf{B}) \setminus \mathbf{Y}.$$

$\mathbf{Y}$ contains those vertices in $\mathit{Border}(\mathbf{B})$ that are not adjacent to $\mathit{Int}(\mathbf{B})$, while $\mathbf{Z}$ has those vertices that are adjacent to $\mathit{Int}(\mathbf{B})$.

By using these subsets we can form the new network.

- In addition to the source and sink, the nodes in the network are the vertices in $\mathbf{S}$ and $\mathit{Border}(\mathbf{B}) = \mathbf{Y} \cup \mathbf{Z}$.
- For each vertex $\mathbf{s} \in \mathbf{S}$, add the directed edge $(\mathbf{source}, \mathbf{s})$ to the network.
- For each vertex $\mathbf{z} \in \mathbf{Z}$, add the directed edge $(\mathbf{z}, \mathbf{sink})$ to the network.
- For $\mathbf{s} \in \mathbf{S}$ and $\mathbf{b} \in \mathbf{Y} \cup \mathbf{Z} = \mathit{Border}(\mathbf{B})$ where $(\mathbf{s}, \mathbf{b})$ is an edge in the original compressed graph, add the directed edge $(\mathbf{s}, \mathbf{b})$ to the network.
- For $\mathbf{y} \in \mathbf{Y}$ and $\mathbf{b} \in \mathbf{Y} \cup \mathbf{Z} = \mathit{Border}(\mathbf{B})$, if $(\mathbf{y}, \mathbf{b})$ is an edge in the original compressed graph, add the directed edge $(\mathbf{y}, \mathbf{b})$ to the network.

- All edges have infinite capacity. For each vertex $\mathbf{s}$ in $\mathbf{S} \cup \mathit{Border}(\mathbf{B})$ we set $\mathit{capacity}(\mathbf{s}) = \mathit{wt}(\mathbf{s})$.

We shall refer to this new network by $\mathcal{N}_w$ (w for wide). Let us first apply the construction on the partition example of Figure 2. We note that the wide separator is subdivided into these three sets:

$$\mathbf{S} = \{3, 7, 8, 9, 13, 19, 25, 26, 32\}, \quad \mathbf{Y} = \{14, 20\}, \quad \text{and} \quad \mathbf{Z} = \{4, 10, 15, 16, 21, 27, 33\}.$$

$\mathcal{N}_w$ is the bottom network of Figure 7. The readers should compare this network with $\mathcal{N}_b$, the original bipartite network $\mathcal{N}_b$, at the top of Figure 7.

We are now ready to establish the important result that this new network $\mathcal{N}_w$ has a max-flow (or min-cut) solution no larger than the one from the bipartite network $\mathcal{N}_b$ using the same wide separator $\mathbf{S} \cup \mathit{Border}(\mathbf{B})$. To prove this result we will construct an intermediate network $\mathcal{N}_m$ by adding the following directed edges into the bipartite network $\mathcal{N}_b$.

- For $\mathbf{y} \in \mathbf{Y}$ and $\mathbf{b} \in \mathbf{Y} \cup \mathbf{Z} = \mathit{Border}(\mathbf{B})$, if $(\mathbf{y}, \mathbf{b})$ is an edge in the original compressed graph, add the directed edge $(\mathbf{y}, \mathbf{b})$ to the network.

$\mathcal{N}_m$ is the middle network in Figure 7 and contains the edges $(14, 20)$, $(20, 14)$, $(14, 21)$, $(14, 15)$, $(20, 15)$, and $(20, 27)$ in addition to those found in $\mathcal{N}_b$.

The following lemma will be used in the next theorem to show that the max-flow values for $\mathcal{N}_b$ and $\mathcal{N}_m$ are identical. It proves that adding an edge connecting two vertices that are both adjacent to the sink does not change the max-flow value.

LEMMA 6.1. *Let x and y be two vertices in a given network $\mathcal{N}_0$, such that both x and y are connected to the sink. Consider the new network $\mathcal{N}_1$ by adding a directed edge (x, y) to $\mathcal{N}_0$. The networks $\mathcal{N}_0$ and $\mathcal{N}_1$ have the same max-flow values.*

Proof. Since the network $\mathcal{N}_1$ has one additional edge than $\mathcal{N}_0$, its max-flow value is at least as large as that of $\mathcal{N}_0$. Consider a flow function f_1 for $\mathcal{N}_1$ that achieves the max-flow value for $\mathcal{N}_1$. If $f_1(x, y) = 0$, there is an equivalent flow function for the network $\mathcal{N}_0$. If $f_1(x, y) > 0$, define the following flow function f_0 for $\mathcal{N}_0$:

$$f_0(x, \mathit{sink}) = f_1(x, \mathit{sink}) + f_1(x, y),$$

$$f_0(y, \mathit{sink}) = f_1(y, \mathit{sink}) - f_1(x, y),$$

$f_0(x, y) = 0$ (there is no directed edge from x to y in $\mathcal{N}_0$), and the f_0 values are the same as the f_1 values for the other vertices. It is easy to see that f_0 is a flow function for $\mathcal{N}_0$ and its flow value is the same as the max-flow value for $\mathcal{N}_1$. □

THEOREM 6.2. *The max-flow values of the networks $\mathcal{N}_b$ and $\mathcal{N}_m$ are the same.*

Proof. First note that the network $\mathcal{N}_m$ is constructed from $\mathcal{N}_b$ by adding a number of directed edges to vertices that are directly linked to the sink. By applying Lemma 6.1 a number of times, we have the result that the networks $\mathcal{N}_b$ and $\mathcal{N}_m$ have the same max-flow values. □

After we delete from $\mathcal{N}_m$ all edges $(\mathbf{y}, \mathbf{sink})$ for $\mathbf{y} \in Y$, (in our example these edges are $(14, \mathbf{sink})$ and $(20, \mathbf{sink})$), we are left with the network $\mathcal{N}_w$. We now show that the max-flow value for $\mathcal{N}_w$ is no larger, and can be smaller, than the max-flow value for $\mathcal{N}_b$. The reach sets from the source and sink are

$$\mathbf{R_{source}} = \{3, 7, 8, 9, 13, 14, 15, 19, 20, 21, 25, 26, 27, 32, 33\},$$
$$\mathbf{R_{sink}} = \{3, 4, 9, 10, 15, 16, 21, 27, 33\},$$

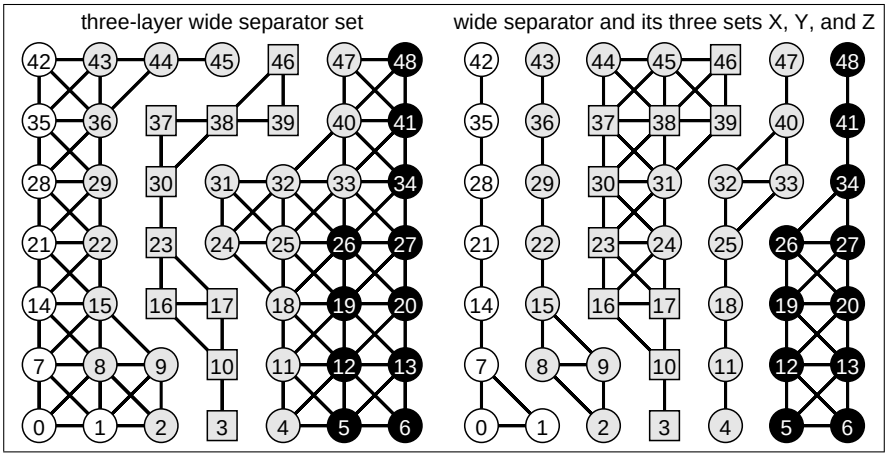


FIG. 9. *Finding a minimal separator using a three-layer network.*

and they both have the same border, and thus give rise to the same min-cut, $\{3, 9, 15, 21, 27, 33\}$, which has six vertices compared with seven vertices for a min-cut of $\mathcal{N}_b$.

THEOREM 6.3. *The max-flow value of the network $\mathcal{N}_w$ is less than or equal to the max-flow value of $\mathcal{N}_b$.*

Proof. Compare the networks $\mathcal{N}_m$ and $\mathcal{N}_w$. The network $\mathcal{N}_w$ can be obtained from $\mathcal{N}_m$ by removing those directed edges $(\mathbf{y}, \mathbf{sink})$ for $\mathbf{y} \in \mathbf{Y}$. Since $\mathcal{N}_w$ is a subnetwork of $\mathcal{N}_m$, the max-flow value of $\mathcal{N}_w$ must be smaller than or equal to that of $\mathcal{N}_m$. □

6.4. Generalization to wider separators. The technique introduced in the last section hinges on the choice of the wide separator $\mathbf{S} \cup \mathit{Border}(\mathbf{B})$. It is easy to generalize this technique for "wider" separators.

Consider a given partition $[\widetilde{\mathbf{S}}, \widetilde{\mathbf{B}}, \widetilde{\mathbf{W}}]$, where the separator set $\widetilde{\mathbf{S}}$ need not be minimal but can be quite large. Subdivide the separator set $\widetilde{\mathbf{S}}$ into three subsets:

$$\mathbf{X} = \mathit{Border}(\widetilde{\mathbf{S}} \cup \widetilde{\mathbf{B}}), \quad \mathbf{Y} = \mathit{Int}(\widetilde{\mathbf{S}}), \quad \text{and} \quad \mathbf{Z} = \mathit{Border}(\widetilde{\mathbf{S}} \cup \widetilde{\mathbf{W}}).$$

A network can be constructed in the same manner as given in the last section by adding edges from the source to vertices in $\mathbf{X}$, from vertices in $\mathbf{Z}$ to the sink, and retaining the underlying edges associated with $\mathbf{Y}$ from the original graph. A max-flow min-cut solution to this network will determine a separator subset of $\widetilde{\mathbf{S}}$ with minimum weight among all such separator subsets.

The wide separator $\mathbf{S} \cup \mathit{Border}(\mathbf{B})$ we have used in our last section can be viewed as having two layers: $\mathbf{S}$ and $\mathit{Border}(\mathbf{B})$. Let us now consider a three-layer separator, given by

$$\widetilde{\mathbf{S}} = \mathit{Border}(\mathbf{W}) \cup \mathbf{S} \cup \mathit{Border}(\mathbf{B}),$$

and solve a flow problem on a three-layer network $\mathcal{N}_3$.

Figure 9 contains an example to illustrate a three-layer separator $\widetilde{\mathbf{S}}$ given by the union of the following three layers:

$$\begin{aligned} \mathit{Border}(\mathbf{W}) &= \{2, 8, 9, 15, 22, 29, 36, 43, 44, 45\}, \\ \mathbf{S} &= \{3, 10, 16, 17, 23, 30, 37, 38, 39, 46\}, \\ \mathit{Border}(\mathbf{B}) &= \{4, 11, 18, 24, 25, 31, 32, 33, 40, 47\}. \end{aligned}$$

TABLE 1
Iteration history for BCSSTK37.

	$\|\mathbf{S}\|$	Imbalance	Reduction in $\|\mathbf{S}\|$	Partition cost γ
Initial two-set partition	1166	1.013	—	2347.2
after block Kernighan–Lin	572	1.118	50.9%	1211.5
1. $\mathbf{S}_I \cup \mathbf{S}_R \mapsto \mathbf{W}$	518	1.062	9.4%	1068.1
2. $\mathbf{S}_I \cup \mathbf{S}_R \mapsto \mathbf{W}$	484	1.038	6.6%	986.4
3. $\mathbf{S}_I \cup \mathbf{S}_R \mapsto \mathbf{W}$	480	1.017	0.8%	968.2
4. $\mathbf{S}_I \mapsto \mathbf{W}$	471	1.001	1.9%	942.5
5. $\mathbf{S}_I \mapsto \mathbf{W}$	460	1.012	2.3%	925.5
6. $\mathbf{S}_I \mapsto \mathbf{W}$	446	1.015	3.0%	898.7
7. $\mathbf{S}_R \mapsto \mathbf{W}$	446	1.013	0.0%	897.8
8. $\mathbf{S}_I \mapsto \mathbf{B}$	438	1.030	1.8%	889.1
9. $\mathbf{S}_I \mapsto \mathbf{B}$	434	1.041	0.9%	885.8
10. $\mathbf{S}_I \mapsto \mathbf{B}$	420	1.051	3.2%	861.4
11. $\mathbf{S}_I \mapsto \mathbf{B}$ rejected	419	1.069	0.2%	867.0

The remaining white vertices form the partition subset $\widetilde{\mathbf{W}}$, while remaining black vertices form $\widetilde{\mathbf{B}}$.

The right grid in Figure 9 shows the decomposition of the wide separator $\widetilde{\mathbf{S}}$ into the three subsets $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. They form the basis on which the network is formed and max-flow min-cut problem is solved. It should be pointed out that often there is more than one min-cut solution. In this example there are three—$\{2, 9, 16, 23, 30, 37, 44\}$, $\{3, 10, 17, 24, 31, 38, 45\}$, and $\{4, 11, 18, 25, 32, 39, 46\}$.

When $\widetilde{\mathbf{S}}$ is even wider, say five or seven layers, the space from which we find a minimal weight separator is large. As the number of layers in $\widetilde{\mathbf{S}}$ increases, the weight of a minimal separator cannot increase. As in our example in Figure 9, there often will be more than one choice of minimal weight separators; we want to choose one that minimizes our partition evaluation function.

7. Experimental results.

7.1. A closer look at two-layer smoothing. In this section, we provide some experimental evidence on improving partitions based on the Dulmage–Mendelsohn decomposition. Table 1 contains a typical iteration history for the algorithm in Figure 1. The sparse matrix BCSSTK37, taken from the Harwell–Boeing collection [7], has 25503 degrees of freedom and 1115474 edges. After compression, we work with the weighted compressed graph with 7093 vertices and 88924 edges.

The partitioning algorithm used is from [2]; readers are referred to it for more details. We first constructed a domain decomposition of the graph—there were 141 domains for this test. The initial partition split the domains into two groups of near equal weight. The separator vertices had weight 1166, and the partition has imbalance of $\max\{|\mathbf{B}|, |\mathbf{W}|\}/\min\{|\mathbf{B}|, |\mathbf{W}|\} = 1.013$. We then applied a block Kernighan–Lin algorithm on the domain-segment graph to reduce the separator size to 572 but with an increase in imbalance to 1.118. The separator at this stage tends to be "locally smooth" when it coincides with the boundary of a domain, but the domains do not generally align themselves to form smooth bisectors of the graph.

We then executed the algorithm in Figure 1. Note that the initial imbalance of 1.118 is rather high. At the first step we evaluate two moves that would reduce the separator size and the size of the larger component, namely, $\mathbf{Z} = \mathbf{S}_I$ and $\mathbf{Z} = \mathbf{S}_I \cup \mathbf{S}_R$. The $\mathbf{S}_I \cup \mathbf{S}_R$ move reduces the partition cost function more. This holds for three moves, as we see both the separator weight and the imbalance decrease together. The

TABLE 2
Statistics for Harwell–Boeing matrices.

	Original		Compressed		MMD		
Matrix	$\|V\|$	$\|E\|$	$\|\mathbf{V}\|$	$\|\mathbf{E}\|$	NZF/10^3	OPS/10^6	CPU
BCSSTK30	28924	2014568	9289	222884	3725	777	1.72
BCSSTK31	35588	1145828	17403	288806	5156	2400	4.70
BCSSTK32	44609	1970092	14821	226974	5147	1048	2.84
BCSSTK33	8738	583166	4344	164284	2654	1301	1.10
BCSSTK35	30237	1419926	6611	65934	2780	406	0.90
BCSSTK36	23052	1120088	4351	37166	1767	626	0.51
BCSSTK37	25503	1115474	7093	88924	2829	548	1.00
BCSSTK39	46772	2042522	10140	81762	7669	2194	1.33
MN12	264002	13115458	51920	569226	40404	24810	12.45
PWT	217918	11634424	41531	483791	63992	49875	7.93

next three moves are $\mathbf{S}_I$ moves, for the balance is close to unity and the $\mathbf{S}_R$ sets are relatively large.

At step 5, note that the move $\mathbf{Z} = \mathbf{S}_I \mapsto \mathbf{W}$ results in a reduction in separator size but an increase in imbalance. After the move the new set $\mathbf{B}_{\mathbf{Z}\mapsto\mathbf{W}}$ is now smaller than $\mathbf{W}_{\mathbf{Z}\mapsto\mathbf{W}}$ and the difference $|\mathbf{W}_{\mathbf{Z}\mapsto\mathbf{W}}| - |\mathbf{B}_{\mathbf{Z}\mapsto\mathbf{W}}|$ is greater than the previous difference $|\mathbf{B}| - |\mathbf{W}|$. At the next step, we maintain the convention that $\mathbf{W}$ is the smaller portion so that the $\mathbf{W}$ in the $\mathbf{S}_I \mapsto \mathbf{W}$ move at step 6 is the $\mathbf{B}_{\mathbf{Z}\mapsto\mathbf{W}}$ from step 5. Again for step 6, there is a reduction in separator size but an increase in imbalance. Step 7 is an instance where the balance is improved with no reduction in separator size. Steps 1–7 are all moves of subsets to the smaller component, so the separator is smoothed in one direction. There is still reduction in the separator to be had by smoothing it against the smaller component, i.e., the larger component absorbs part of the separator, as we see in steps 8–11. The separator weight decreases by 5.9% during steps 8–10 while the imbalance increases from 1.013 to 1.051. At step 11 there is still a possible reduction in separator weight, where $|\mathbf{Z}| = 106$ and $|Adj_H(\mathbf{Z})| = 105$. Making this move would increase the partition cost function, so the algorithm terminates.

7.2. Comparing two-layer and three-layer smoothers. We have tested the various partition improvement techniques described in this paper on a collection of test matrix problems. Table 2 contains the description of 10 sparse matrix problems from the Harwell–Boeing collection [7].

Table 3 presents statistics for finding a top-level separator for the three algorithms. The cost is $|S|\left(1+\alpha\frac{\max(|B|,|W|)}{\min(|B|,|W|)}\right)$, where the penalty multiplier $\alpha = 1$. The median cost value for 25 runs is found in the table—for each run the matrix was randomly permuted. The initial partition is obtained from domain decomposition followed by the block Kernighan–Lin scheme in [2] as discussed in the last section.

The three algorithms tested are labeled $\mathcal{N}_b$, $\mathcal{N}_w$, and $\mathcal{N}_3$, respectively, in the table. Column $\mathcal{N}_b$ has statistics for the partition improvement algorithm in Figure 1 using the Dulmage–Mendelsohn decomposition, i.e., it solves the max-flow problem defined on the bipartite network $\mathcal{N}_b$. Column $\mathcal{N}_w$ contains results for the partition improvement algorithm in Figure 1 using the two-layer wide network $\mathcal{N}_w$. These two algorithms iterate until no improvement can be made. Inside the loop, they make a first attempt to improve the partition based on a two-layer separator $\mathbf{S} \cup \mathit{Border}(\mathbf{B})$ using the current separator $\mathbf{S}$ and the larger portion $\mathbf{B}$. If there is no improvement on

TABLE 3
Top-level separators, median cost of 25 *runs.*

	Using $\mathcal{N}_b$			Using $\mathcal{N}_w$			Using $\mathcal{N}_3$		
Matrix	Cost	$\|S\|$	Balance	Cost	$\|S\|$	Balance	Cost	$\|S\|$	Balance
BCSSTK30	467	223	1.095	421	209	1.012	421	209	1.012
BCSSTK31	707	353	1.001	679	339	1.003	680	332	1.049
BCSSTK32	791	355	1.228	717	322	1.226	711	271	1.624
BCSSTK33	847	421	1.012	847	421	1.012	847	421	1.012
BCSSTK35	344	162	1.121	306	144	1.128	307	96	2.194
BCSSTK36	715	357	1.002	644	325	1.043	662	331	1.000
BCSSTK37	894	440	1.031	889	437	1.033	889	437	1.033
BCSSTK39	451	225	1.003	451	225	1.003	451	225	1.003
MN12	1736	861	1.017	1662	815	1.039	1609	791	1.034
PWT	1441	720	1.001	1441	720	1.001	1442	720	1.003

TABLE 4
Nested dissection compared to multiple minimum degree; a value greater than one means that nested dissection generates more factor entries, operations, or CPU than minimum degree.

	Factor entries			Factor ops			Ordering cpu		
Matrix	$\mathcal{N}_b$	$\mathcal{N}_w$	$\mathcal{N}_3$	$\mathcal{N}_b$	$\mathcal{N}_w$	$\mathcal{N}_3$	$\mathcal{N}_b$	$\mathcal{N}_w$	$\mathcal{N}_3$
BCSSTK30	1.24	1.11	1.13	1.88	1.42	1.46	4.86	5.16	6.07
BCSSTK31	0.89	0.84	0.84	0.58	0.52	0.50	3.21	3.20	3.48
BCSSTK32	1.12	1.09	1.07	1.48	1.38	1.33	4.40	3.96	4.19
BCSSTK33	0.86	0.83	0.80	0.71	0.65	0.57	4.86	4.74	7.21
BCSSTK35	1.15	1.11	1.09	1.55	1.41	1.36	4.20	4.13	4.20
BCSSTK36	1.13	1.07	1.07	1.42	1.25	1.25	4.47	4.47	4.47
BCSSTK37	1.09	1.07	1.06	1.36	1.35	1.30	4.24	4.29	4.42
BCSSTK39	0.94	0.94	0.94	0.95	0.94	0.94	4.11	4.11	4.05
MN12	1.08	1.00	0.97	1.07	0.92	0.82	3.53	3.56	3.57
PWT	0.74	0.74	0.74	0.47	0.47	0.46	4.26	4.22	4.36

this attempt, it will then try the two-layer separator $\mathbf{S} \cup \mathit{Border}(\mathbf{W})$ with the smaller portion $\mathbf{W}$.

The algorithm associated with $\mathcal{N}_3$ is also iterative in nature. It is simpler since it tries to improve the partition using the three-layer set $\mathbf{S} \cup \mathit{Border}(\mathbf{B}) \cup \mathit{Border}(\mathbf{W})$. It continues until no improvement can be obtained. Our experience shows that the algorithm for $\mathcal{N}_3$ typically requires half the number of steps or less when compared to the first two algorithms. But, of course, it takes more time at each step since it is solving a larger network problem. We see that often using the network $\mathcal{N}_w$ gives sizable partition improvement over the network $\mathcal{N}_b$. Using the three-layer network sometimes gives additional but small improvement.

We have also used the three partition improvement algorithms to find separators in the context of finding fill-reducing sparse matrix orderings. Tables 4 and 5 contain statistics of nested dissection orderings and multisection orderings [3] using the three partition improvement schemes. The statistics are scaled by results from the multiple minimum degree ordering. Each result in the tables comes from the run that generated the median factor operations in 25 runs.

We have experimented with using a network with five layers, seven layers, and more to improve separators. Any improvement is usually modest while the run times for the orderings increase dramatically as the time to solve the max-flow problems for the larger networks takes a larger portion of the ordering time.

Wide separators have a disadvantage for the min-cuts may be spread across the wide separator. Consider an example where we start with a partition that has good

TABLE 5
Multisection compared with multiple minimum degree; a value greater than one means that multisection generates more factor entries, operations, or CPU than minimum degree.

	Factor entries			Factor ops			Ordering cpu		
Matrix	$\mathcal{N}_b$	$\mathcal{N}_w$	$\mathcal{N}_3$	$\mathcal{N}_b$	$\mathcal{N}_w$	$\mathcal{N}_3$	$\mathcal{N}_b$	$\mathcal{N}_w$	$\mathcal{N}_3$
BCSSTK30	1.09	1.01	1.04	1.30	1.08	1.15	4.87	5.16	6.07
BCSSTK31	0.90	0.86	0.85	0.61	0.58	0.55	3.19	3.22	3.49
BCSSTK32	0.97	0.95	0.94	0.90	0.85	0.84	4.04	3.96	4.19
BCSSTK33	0.81	0.79	0.79	0.61	0.57	0.57	4.86	4.74	7.20
BCSSTK35	1.03	1.00	0.99	1.06	1.01	0.97	4.20	4.12	4.19
BCSSTK36	0.96	0.94	0.94	0.85	0.82	0.82	4.49	4.48	4.47
BCSSTK37	0.95	0.94	0.93	0.87	0.85	0.84	4.24	4.28	4.41
BCSSTK39	0.89	0.89	0.90	0.77	0.78	0.79	4.12	4.11	4.05
MN12	1.00	0.94	0.93	0.88	0.77	0.75	3.53	3.58	3.56
PWT	0.79	0.79	0.79	0.59	0.60	0.59	4.25	4.21	4.36

balance. When we use a very wide separator (say seven levels) to form a network, a min-cut may lie far to one side or the other of the "thin" separator. Though the separator induced by the min-cut might be smaller than the present separator, the partition that would result may have a larger cost due to an increased imbalance, and so the new partition would not be accepted. There is one min-cut closest to the source and one closest to the sink (the two may be identical), and neither might result in a better partition. We are not primarily interested in finding the minimal weight separator—we want a partition whose cost is minimal. To this end we are exploring ways to modify the network such that the min-cut determines a partition with minimal cost.

8. Concluding remarks. In this paper, we have presented a detailed exposition of the Dulmage–Mendelsohn decomposition of bipartite graphs in the context of improving bisector-based partitions. In the literature, this decomposition has been used to obtain a vertex separator from an edge separator, and in iteratively improving a vertex separator. We have also used the decomposition to improve the balance of a partition.

Another contribution of this paper is the extension of the Dulmage–Mendelsohn decomposition to compressed graphs, a special type of weighted graphs that occur naturally and frequently in practice. For such graphs, we have related the decomposition with the well-known maximum flow network problem. Finding a separator of minimum cover based on the Dulmage–Mendelsohn decomposition is the same as obtaining a min-cut of a bipartite network problem. We have also introduced an enhancement by solving a slightly modified network problem, the solution of which will often yield a smaller separator.

We have provided experimental results to demonstrate the viability of the approaches to improve bisectors and partitions. These results should be viewed as additional evidence to those included in our earlier paper [2]. We recommend this smoothing step using graph matching or network max-flow min-cut as a standard final process on all dissection-based ordering codes. Indeed, such smoother codes are present in the recently developed software, such as the new CHACO code [17] by Hendrickson and Rothberg, and the IBM Watson Graph Partitioning code WGPP [13] by Gupta.

Max-flow techniques have potential applications in other contexts, particularly to find separators of coarse graphs used in multilevel algorithms [13], [17], [20] or the domain/segments graphs from a domain-decomposition approach [2]. While we have

concentrated on "thin" networks, where the distance from the source to the sink is small, in principal one can attack much wider separators, perhaps containing all of a graph save for a source and sink vertex. While this would be prohibitively expensive for a large graph, it could be profitably used for a coarse graph or domain/segment graph. The drawback is that a min-cut might naturally lie very close to the source or sink and thus induce a poorly balanced partition. By increasing the weight of vertices close to the source or sink one can force the min-cut to split the graph into two more equally sized pieces [10].

Acknowledgments. We would like to thank Matt Berge of Boeing Information and Support Services for several enlightening conversations on network flow and Stan Eisenstat of Yale University for many insightful comments on an earlier draft. We owe a great debt to John Gilbert of Xerox PARC for his notes on bipartite graphs and the Dulmage–Mendelsohn decomposition.

REFERENCES

[1] C. Ashcraft, *Compressed graphs and the minimum degree algorithm*, SIAM J. Sci. Comput., 16 (1995), pp. 1404–1411.
[2] C. Ashcraft and J. Liu, *Using domain decomposition to find graph bisectors*, BIT, 37 (1997), pp. 506–534.
[3] C. Ashcraft and J. Liu, *Robust ordering of sparse matrices using multisection*, SIAM J. Matrix Anal. Appl., to appear.
[4] C. Ashcraft and J. Liu, *Generalized nested dissection: some recent progress*, in Proc. Fifth SIAM Conference on Applied Linear Algebra, SIAM, Philadelphia, PA, 1994, pp. 130–134.
[5] T. Bui and C. Jones, *A heuristic for reducing fill-in in sparse matrix factorization*, in Proc. Sixth SIAM Conference on Parallel Processing, SIAM, Philadelphia, PA, 1993, pp. 445–452.
[6] A. C. Damhaug, *Sparse Solution of Finite Element Equations*, Ph.D. thesis, The Norwegian Institute of Technology, 1992.
[7] I. Duff, R. Grimes, and J. Lewis, *Sparse matrix test problems*, ACM Trans. Math Software, 15 (1989), pp. 1–14.
[8] A. Dulmage and N. Mendelsohn, *Coverings of bipartite graphs*, Canad. J. Math, 10 (1958), pp. 517–534.
[9] J. Edmonds and R. M. Karp, *Theoretical improvements in algorithmic efficiency for network flow problem*, J. ACM, 19 (1972), pp. 248–264.
[10] S. C. Eisenstat, *private communication*, Yale University, New Haven, CT, 1996.
[11] L. R. Ford and D. R. Fulkerson, *Flows in Networks*, Princeton University Press, Princeton, NJ, 1962.
[12] D. R. Fulkerson, *Flow networks and combinatorial operations research*, Amer. Math. Monthly, 73 (1966), pp. 115–136.
[13] A. Gupta, *WGPP: Watson Graph Partitioning and Sparse Matrix Ordering Package: Users Manual*, Tech. report RC 20453 (90427), IBM T. J. Watson Research Center, Yorktown Heights, NY, 1996.
[14] P. Hall, *On representatives of subsets*, J. London Math. Soc., 10 (1935), pp. 26–30.
[15] B. Hendrickson and R. Leland, *The Chaco User's Guide*, Tech. report SAND93-2339, Sandia National Laboratories, Albuquerque, NM, 1993.
[16] B. Hendrickson and R. Leland, *An improved spectral graph partitioning algorithm for mapping parallel computations*, SIAM J. Sci. Comput., 16 (1995), pp. 452–469.
[17] B. Hendrickson and E. Rothberg, *Improving the runtime and quality of nested dissection ordering*, SIAM J. Sci. Comput., to appear.
[18] A. Itai and Y. Shiloach, *Maximum flow in planar networks*, SIAM J. Comput., 8 (1979), pp. 135–150.
[19] G. Karypis and V. Kumar, *A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs*, Tech. report TR 95-035, Department of Computer Science, University of Minnesota, Minneapolis, MN, 1995.
[20] G. Karypis and V. Kumar, *Metis: Unstructured Graph Partitioning and Sparse Matrix Ordering System*, Tech. report TR 97-061, Department of Computer Science, University of Minnesota, Minneapolis, MN, 1995.

[21] B. W. Kernighan and S. Lin, *An efficient heuristic procedure for partitioning graphs*, Bell System Tech. J., 49 (1970), pp. 291–307.
[22] E. L. Lawler, *Combinatorial Optimization: Networks and Matriods*, Holt, Rinehart and Winston, New York, 1967.
[23] C. E. Leiserson and J. G. Lewis, *Ordering for parallel sparse symmetric factorization*, in Parallel Processing for Scientific Computing, SIAM, Philadelphia, PA, 1989, pp. 27–31.
[24] J. W. H. Liu, *A graph partitioning algorithm by node separators*, ACM Trans. Math Software, 15 (1989), pp. 198–219.
[25] A. Pothen, *Sparse Null Bases and Marriage Theorems*, Ph.D. thesis, Cornell University, Ithaca, NY, 1984.
[26] A. Pothen and C. Fan, *Computing the block triangular form of a sparse matrix*, ACM Trans. Math Software, 16 (1990), pp. 303–324.
[27] A. Pothen, H. Simon, and K. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.
[28] R. E. Tarjan, *Data Structures and Network Algorithms*, SIAM, Philadelphia, PA, 1983.
[29] J. D. Ullman, *Computational Aspects of VLSI*, Computer Science Press, Rockville, MD, 1984.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 355–364, April 1998

004

SIGN CONTROLLABILITY: SIGN PATTERNS THAT REQUIRE COMPLETE CONTROLLABILITY*

MICHAEL J. TSATSOMEROS†

Abstract. We apply tools from the theory of sign-solvable systems and use the directed graph of a matrix in order to obtain sufficient conditions for a linear control system (A, B) to be completely controllable solely due to the sign patterns of the coefficient matrices A and B. We show that such conditions are necessary and sufficient for a particular class of linear control systems. We also consider an alternative approach to controllability, based on a reformulation of the classical condition (that the controllability matrix is of full rank) and obtain equivalent conditions for the general case.

Key words. control system, sign pattern, signing, L-matrix, directed graph, aligned vertices, balancing chain

AMS subject classifications. 15A99, 93B05

PII. S0895479896300346

1. Introduction.

In linear control theory, the basic concepts of controllability (and observability) are intimately related to the image of a matrix of the form

$$\mathcal{C} \;=\; [B \;\; AB \;\; \ldots \;\; A^{n-1}B] \in \mathbf{R}^{n,\ nm},$$

where $A \in \mathbf{R}^{n,\ n}$ and $B \in \mathbf{R}^{n,\ m}$. Specifically, a control system of the form

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t)$$

is completely controllable if and only if rank $\mathcal{C} = n$. As the matrices A, B comprise system parameters prone to measurement errors, it is desirable to determine whether rank $\mathcal{C} = n$ based on combinatorial and qualitative information about A and B (e.g., their directed graphs and the signs of their entries). Such qualitative approaches to controllability have been undertaken, for example, by Lin [6], Mayeda and Yamada [9], Murota [8], Johnson, Mehrmann, and Olesky [4], and Olesky, Tsatsomeros, and van den Driessche [10].

In the present work, we shall consider the following: assume that the sign patterns (namely, the location of the positive, negative, and zero entries) of A and B are known. *When can we conclude that* rank $\mathcal{C} = n$*, based solely on the sign patterns and regardless of the magnitudes of the nonzero entries of* A *and* B*?* The study of this question was initiated in [4], where A was assumed to have nonnegative entries and B was assumed to be a column vector with positive entries.

Our qualitative approach to controllability will be based on extending and combining techniques used in the study of zero/nonzero patterns that *allow* or *require* complete controllability (see [6, 9, 10]) with notions related to the analysis of sign patterns and sign-solvable linear systems. We will find sufficient conditions for complete controllability for general sign patterns (Theorem 3.2), and we will identify a

*Received by the editors March 11, 1996; accepted for publication (in revised form) by V. Mehrmann February 6, 1997. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

http://www.siam.org/journals/simax/19-2/30034.html

†Department of Mathematics and Statistics, University of Regina, Regina, Saskatchewan S4S 0A2, Canada (tsat@math.uregina.ca).

class of linear control systems (Definition 2.2) for which these conditions are necessary and sufficient (Theorem 3.7). In section 4, we will consider a simple technical recasting of the classical controllability condition that rank $\mathcal{C} = n$ in order to provide an alternative answer to the question posed above.

2. Preliminaries. In this section, we present some of the notation, terminology, and basic facts necessary to state and prove the main results in the following sections. In the remainder we let $\langle k \rangle = \{1, 2, \ldots, k\}$ for any positive integer k; $\mid \alpha \mid$ denote the cardinality of a set α; $\operatorname{sgn}(a)$ be 0, 1, or -1, when a is zero, positive, or negative, respectively; $\operatorname{Re}(x)$ denote the real part of a complex vector x; $\operatorname{diag}(A_1, A_2, \ldots, A_k)$ be a block diagonal matrix whose diagonal blocks are the square matrices $A_1, A_2, \ldots, A_k$; $X[\alpha \mid \beta]$ denote the submatrix of $X \in \mathbf{R}^{s,\,t}$ whose rows and columns are indexed by the sets $\alpha \subseteq \langle s \rangle$ and $\beta \subseteq \langle t \rangle$, respectively; $X[\alpha] = X[\alpha \mid \alpha]$; and e denote an all 1s column vector of appropriate size.

We let $\Gamma = (V, E)$ denote a *directed graph* with vertex set V and directed edge set E consisting of ordered pairs (i, j) of vertices. 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 $i = 1, \ldots, t-1$.

The *directed graph of* $X = (x_{ij}) \in \mathbf{R}^{s,\,t}$, $s \leq t$, denoted by $\Gamma = \mathcal{D}(X) = (V, E)$, has $V = \langle t \rangle$ and $E = \{(i, j) \mid x_{ij} \neq 0\}$. Extending the terminology in [6] and [9], when $s < t$ we say that $\mathcal{D}(X)$ is *accessible* if for every $j \in \langle s \rangle$ there exists $k \in \langle t \rangle \setminus \langle s \rangle$, such that there is a path from j to k in Γ. Also, for every $\alpha \subseteq \langle s \rangle$ we denote the *adjacency set* of α by

$$\mathcal{R}(\alpha) = \{j \in \langle t \rangle \mid (i, j) \in E \ \text{ for some } \ i \in \alpha\}.$$

Notice that if there exists an $\alpha \subseteq \langle s \rangle$ such that $\alpha \subseteq \mathcal{R}(\alpha)$ and if Γ is accessible, then $\mathcal{R}(\alpha) \cap (\langle t \rangle \setminus \langle s \rangle) \neq \emptyset$ and hence $\alpha \subset \mathcal{R}(\alpha)$.

In keeping with the notation and terminology of Brualdi and Shader [1] (which is our comprehensive reference on sign patterns), we define the following.

The *sign pattern* of $X \in \mathbf{R}^{s,\,t}$ is the $(0, 1, -1)$-matrix obtained from X when zero, positive, and negative entries are replaced by 0, 1, and -1, respectively. The matrix X determines the *qualitative class* $Q(X)$ of all matrices with the same sign pattern as X. We will write $\hat{X} \in Q(X)$ for any matrix $\hat{X}$ having the same sign pattern as X.

A *signing* is a nonzero square diagonal sign pattern. A real vector is called *balanced* if it is the zero vector, or if it has at least one negative and at least one positive entry. A real vector is referred to as *unisigned* if it is not balanced. If a unisigned vector has nonnegative (respectively, nonpositive) entries, we refer to it as *of positive (*respectively, *negative) type*. We denote the signings S such that all the columns of SX are balanced by $\mathcal{B}(X)$.

The matrix X is called an *L-matrix* provided that every matrix in $Q(X)$ has linearly independent rows. It is well known (see [1, Theorem 2.1.1]) that $X \in \mathbf{R}^{s,\,t}$ is an L-matrix if and only if $\mathcal{B}(X) = \emptyset$. Next we introduce the notion of aligned vertices in the directed graph of a matrix.

DEFINITION 2.1. *Let $X \in \mathbf{R}^{s,\,t}$, $s \leq t$, and let $\alpha_1 \subseteq \langle s \rangle, \alpha_2 \subseteq \langle t \rangle$ be two nonempty and disjoint sets. We call* α_1 aligned relative to α_2 *if there exists a signing $S \in \mathcal{B}(X[\alpha_1 \mid \alpha_2])$ such that the unisigned columns of $SX[\alpha_1]$ (if any exist) are only of one type (either only positive type or only negative type). When $\mathcal{B}(X[\alpha_1 \mid \alpha_2]) = \emptyset$, α_1 is by definition not aligned relative to α_2.*

In other words, $\alpha_1 \subseteq \langle s \rangle$ is aligned relative to a disjoint set $\alpha_2 \subseteq \langle t \rangle$ if there exists a signing of the rows of $X[\alpha_1 \mid \langle t \rangle]$ such that the columns of $X[\alpha_1 \mid \alpha_2]$ become

balanced and the columns of $X[\alpha_1]$ become either balanced or unisigned of only one type.

Consider now a linear control system of the form

$$\frac{d}{dt}x(t) = Ax(t) + Bu(t), \quad t \geq 0, \tag{2.1}$$

where $A \in \mathbf{R}^{n,\,n}$ and $B \in \mathbf{R}^{n,\,m}$, and where $u(t) \in \mathbf{R}^m$ represents an unconstrained, piecewise continuous control input. We denote the system in (2.1) by (A, B). It is known that the output (viz. solution) $x(t)$ of (2.1) emanating from any initial point in $\mathbf{R}^n$ is controllable (by an appropriate choice of $u(t)$) to any terminal point in $\mathbf{R}^n$ in finite time if and only if

$$\operatorname{rank} \mathcal{C} = n, \tag{2.2}$$

where $\mathcal{C} = [B \;\; AB \;\; \ldots \;\; A^{n-1}B] \in \mathbf{R}^{n,\,nm}$ is the *controllability matrix* associated with (A, B). When (2.2) holds, we call (A, B) *completely controllable*. It follows easily that if $X \in \mathbf{R}^{n,\,n}$ is nonsingular, then (A, B) is completely controllable if and only if (XAX^{-1}, XB) is completely controllable, or if and only if $(-A, B)$ is completely controllable.

As with many questions arising in the study of sign patterns, the presence of implicit relations among the entries of the matrix in question can complicate the qualitative analysis significantly. In the case of the controllability matrix $\mathcal{C}$ this difficulty is evident because of the presence of the products of powers of A with B. For this reason, it is useful to consider a condition known to be equivalent to $\operatorname{rank} \mathcal{C} = n$ (see, e.g., Theorem 4.3.3 in Lancaster and Rodman [7]), namely,

$$\operatorname{rank}[A - \lambda I \;\; B] = n \;\text{ for all }\; \lambda \in \mathbf{C}. \tag{2.3}$$

The compromise in dealing with the latter condition, rather than $\mathcal{C}$, is the introduction of the complex parameter λ.

Given a linear control system (A, B), we consider the qualitative class consisting of all linear control systems $(\hat{A}, \hat{B})$ such that $\hat{A} \in Q(A)$, $\hat{B} \in Q(B)$. In this paper, we say that (A, B) is *sign controllable*[1] if $(\hat{A}, \hat{B})$ is completely controllable for all $\hat{A} \in Q(A)$ and all $\hat{B} \in Q(B)$.

Next we introduce a classification of control systems (A, B) based on the directed graph of $[A\ B]$ and the signs of the diagonal entries of A.

DEFINITION 2.2. *Let $A \in \mathbf{R}^{n,\,n}$, $B \in \mathbf{R}^{n,\,m}$, $T = [A\ B]$, $\Gamma = \mathcal{D}(T)$. We call (A, B) a* strict linear control system *if*

(a) *the diagonal entries of A are nonzero and have the same sign, and*

(b) *for all $\alpha \subseteq \langle n \rangle$ such that $\alpha \subset \mathcal{R}(\alpha)$ in Γ, either $T[\alpha \mid \mathcal{R}(\alpha)\backslash\alpha]$ is an L-matrix or $\mathcal{B}(T[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha])$ contains a nonsingular signing.*

In the next section, we will find sufficient conditions for sign controllability and we will show that these conditions are necessary and sufficient for sign controllability of a strict linear control system.

3. Conditions for sign controllability. First we mention a necessary condition for complete controllability (that is observed in [6] as a necessary condition for *structural controllability*).

[1] We caution the reader that the term *sign controllability* has also been used in the literature to describe a different property of the controllability matrix (see [3, 7]).

LEMMA 3.1. *Let $A \in \mathbf{R}^{n,\,n}$, $B \in \mathbf{R}^{n,\,m}$, and suppose that (A, B) is completely controllable. Then $\Gamma = \mathcal{D}([A\ B])$ is accessible.*

Proof. Suppose that Γ is not accessible. Then there exists $j \in \langle n \rangle$ such that there is no path from j to k in Γ for every $k \in \langle n+m \rangle \setminus \langle n \rangle$. Let $\alpha \subseteq \langle n \rangle$ be the set consisting of j and all vertices of $\mathcal{D}(A)$ that lie on a path emanating from j. It follows that there is no path from ℓ to k in Γ for every $\ell \in \alpha$ and every $k \in \langle n+m \rangle \setminus \langle n \rangle$. Moreover, letting α^c be the complement of α in $\langle n \rangle$, there exists a permutation matrix P such that

$$PAP^T = \begin{bmatrix} A[\alpha] & 0 \\ A[\alpha^c \mid \alpha] & A[\alpha^c] \end{bmatrix} \quad \text{and} \quad PB = \begin{bmatrix} 0 \\ B[\alpha^c \mid \langle m \rangle] \end{bmatrix}.$$

So, if $x = [\hat{x}^T\ 0]^T \in \mathbf{R}^n$, where $\hat{x}$ is a left eigenvector of $A[\alpha]$ corresponding to an eigenvalue λ, then $x^T[PAP^T - \lambda I \quad PB] = 0$, showing that (PAP^T, PB) (and hence (A, B)) is not completely controllable. ∎

From condition (2.3) for $\lambda = 0$ and the above lemma, we have that $[A\ B]$ being an L-matrix and the directed graph of $[A\ B]$ being accessible are two necessary conditions for sign controllability of (A, B). We continue by showing that these two conditions, together with some additional conditions on the directed graph of $[A\ B]$, are also sufficient for sign controllability of (A, B).

THEOREM 3.2. *Let $A \in \mathbf{R}^{n,\,n}$, $B \in \mathbf{R}^{n,\,m}$, and $\Gamma = \mathcal{D}([A\ B]) = (V, E)$. Suppose that*

(1) *Γ is accessible,*

(2) *$[A\ B]$ is an L-matrix, and*

(3) *for all $\alpha \subseteq \langle n \rangle$ satisfying $\alpha \subset \mathcal{R}(\alpha)$ in Γ, either there exists $j \in \mathcal{R}(\alpha) \setminus \alpha$ and exactly one $i \in \alpha$ such that $(i, j) \in E$, or α is not aligned relative to $\mathcal{R}(\alpha) \setminus \alpha$.*

Then the linear control system (A, B) is sign controllable.

Proof. Suppose (A, B) is not completely controllable and that (1) and (2) hold. It is enough to show that (3) is not true. By condition (2.3) and because $[A\ B]$ is an L-matrix, there exists $\lambda \in \mathbf{C} \setminus \{0\}$ and $x \in \mathbf{C}^n \setminus \{0\}$ such that

$$x^T[A\ B] = [\lambda x^T\ 0]. \tag{3.1}$$

Without loss of generality, assume that $x = (x_1, x_2, \ldots, x_k, 0, \ldots 0)^T$, $x_i \neq 0$ for $i = 1, 2, \ldots, k$ (otherwise we can work with (PAP^T, PB) for some permutation matrix P). Also without loss of generality assume that $\mathrm{Re}(x) \neq 0$ (otherwise we can replace x in our arguments by $\sqrt{-1}x$). Observe that $\mathrm{Re}(\lambda x^T) \neq 0$ or else, by (3.1), $\mathrm{Re}(x^T)[A\ B] = 0$ and (2) is contradicted. Consider an invertible signing $S = \mathrm{diag}(s_1, s_2, \ldots, s_n)$ so that $\mathrm{Re}(\lambda x^T S) \geq 0$ (entrywise). On letting $T = [SAS\ SB] = (t_{ij})$, we have from (3.1) that

$$x^T S[SAS\ SB] = [\lambda x^T S\ 0], \tag{3.2}$$

namely,

$$\sum_{i=1}^{k} x_i s_i t_{ij} = \lambda x_j s_j \neq 0 \quad (j = 1, 2, \ldots, k). \tag{3.3}$$

Now take $\alpha = \langle k \rangle \subseteq \langle n \rangle$ and let α^c be its complement in $\langle n+m \rangle$. From (3.3) we can conclude that every column of $T[\alpha]$ contains at least one nonzero entry. Hence for

every $j \in \alpha$ there exists $i \in \alpha$ such that $(i,j) \in E$. This means that $\alpha \subseteq \mathcal{R}(\alpha)$. Since Γ is assumed accessible, we have that $\alpha \subset \mathcal{R}(\alpha)$. We also have from (3.2) that

$$\sum_{i=1}^{k} x_i s_i t_{ij} = 0 \quad (j = k+1, k+2, \ldots, n+m), \tag{3.4}$$

which implies that every column of $T[\alpha \mid \alpha^c]$ has either no or at least two nonzero entries. Hence for every $j \in \mathcal{R}(\alpha) \setminus \alpha \subseteq \alpha^c$ there are at least two vertices $i \in \alpha$ such that $(i,j) \in E$. As a consequence, to show that condition (3) is violated it remains to argue that α is aligned relative to $\mathcal{R}(\alpha) \setminus \alpha$. From (3.3) and (3.4) we get, respectively, that

$$\sum_{i=1}^{k} \mathrm{Re}(x_i) s_i t_{ij} = \mathrm{Re}(\lambda x_j s_j) \geq 0 \quad (j = 1, 2, \ldots, k) \tag{3.5}$$

and

$$\sum_{i=1}^{k} \mathrm{Re}(x_i) s_i t_{ij} = 0 \quad (j = k+1, k+2, \ldots, n+m). \tag{3.6}$$

Equations (3.5) and (3.6) have the following interpretation: if we consider the signing

$$\hat{S} = \mathrm{diag}(\mathrm{sgn}(\mathrm{Re}(x_1))s_1, \mathrm{sgn}(\mathrm{Re}(x_2))s_2, \ldots, \mathrm{sgn}(\mathrm{Re}(x_k))s_k),$$

then the columns of $\hat{S}T[\alpha \mid \alpha^c]$, and in particular the columns of $\hat{S}T[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha]$, are balanced, while all unisigned columns of $\hat{S}T[\alpha]$ are of positive type. Hence α is aligned relative to $\mathcal{R}(\alpha) \setminus \alpha$ in Γ. □

We continue with some examples in order to illustrate the use of Theorem 3.2 and various situations that arise.

Example 3.3. Let

$$A = \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Notice that the directed graph of $[A\ B]$ is accessible and that $[A\ B]$ is an L-matrix because $\det\hat{A} < 0$ for all $\hat{A} \in Q(A)$ (i.e., A is a sign nonsingular matrix; see [1]). Regarding condition (3) of Theorem 3.2, we find that $\alpha_i \subset \mathcal{R}(\alpha_i)$, $i = 1, 2, 3, 4$, where $\alpha_1 = \{3\}$, $\alpha_2 = \{1, 2\}$, $\alpha_3 = \{1, 3\}$, and $\alpha_4 = \langle 3 \rangle$. We also have that

$$\mathcal{R}(\alpha_1) \setminus \alpha_1 = \{1\}, \quad \mathcal{R}(\alpha_2) \setminus \alpha_2 = \{3, 4\}, \quad \mathcal{R}(\alpha_3) \setminus \alpha_3 = \{2, 4\}, \quad \text{and} \quad \mathcal{R}(\alpha_4) \setminus \alpha_4 = \{4\}.$$

In all four cases, the first part of condition (3) is satisfied with the edge from α_i to $\mathcal{R}(\alpha_i) \setminus \alpha_i$, $i = 1, 2, 3, 4$, being $(3, 1)$, $(1, 4)$, $(1, 4)$, and $(1, 4)$, respectively. So by Theorem 3.2, (A, B) is sign controllable. We comment that, in the language of [10], the pair of zero/nonzero patterns $(\mathbf{A}, \mathbf{B})$ associated with (A, B) is not *qualitatively controllable* (see ([10, Theorem 2.2]).

Example 3.4. Let

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & -1 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}.$$

The directed graph of $T = [A\ B]$ is accessible and T is an L-matrix. Regarding condition (3) of Theorem 3.2, we find that with one exception, for all $\alpha \subseteq \langle 3\rangle$ for which $\alpha \subset \mathcal{R}(\alpha)$, there is exactly one edge from α to some $j \in \mathcal{R}(\alpha) \setminus \alpha$. The only exception is $\hat{\alpha} = \langle 3\rangle$ for which $\mathcal{R}(\hat{\alpha}) \setminus \hat{\alpha} = \{4, 5\}$. Notice that every $S \in \mathcal{B}(T[\hat{\alpha} \mid \mathcal{R}(\hat{\alpha}) \setminus \hat{\alpha}])$ has its first two diagonal entries zero and the third diagonal entry nonzero. But then the last two columns of $ST[\hat{\alpha}]$ are unisigned of opposite type. Hence, by Theorem 3.2, (A, B) is sign controllable.

We continue with a result on sign controllability, which will lead to a characterization of strict sign controllable systems.

PROPOSITION 3.5. *Let $A \in \mathbf{R}^{n,\,n}$, $B \in \mathbf{R}^{n,\,m}$, $T = [A\ B]$, and $\Gamma = \mathcal{D}(T)$. Assume that there exists $\alpha \subseteq \langle n\rangle$ with $\alpha \subset \mathcal{R}(\alpha)$ in Γ such that $\mathcal{B}(T[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha])$ contains a nonsingular signing S. Also assume that the unisigned columns of $ST[\alpha]S$ (if any exist) are only of one type. Then (A, B) is not sign controllable.*

Proof. Let S be as prescribed above and $\Gamma = (V, E)$. Since (A, B) is completely controllable if and only if $(-A, B)$ is, we will assume, without loss of generality, that the unisigned columns of $ST[\alpha]S$ (if any exist) are all of positive type. Since there is no $i \in \alpha$ and $j \notin \mathcal{R}(\alpha)$ such that $(i, j) \in E$, $T[\alpha \mid \langle n + m\rangle \setminus \mathcal{R}(\alpha)] = 0$. Also, since $\alpha \subset \mathcal{R}(\alpha)$, every column of $T[\alpha]$ contains a nonzero entry. Letting $\hat{S} \in \mathbf{R}^{n,\,n}$ be a nonsingular signing such that $\hat{S}[\alpha] = S$ and considering $\hat{T} = [\hat{S}A\hat{S}\ \hat{S}B]$, we have that

(1) every column of $\hat{T}[\alpha]$ contains a positive entry,
(2) every column of $\hat{T}[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha]$ is balanced, and
(3) every column of $\hat{T}[\alpha \mid \langle n + m\rangle \setminus \mathcal{R}(\alpha)]$ is zero.

Therefore, by (1)–(3) above, we can assume that the nonzero entries of A and B have been chosen so that the entries of each column of $\hat{T}[\alpha]$ add up to one, and the entries of each column of $\hat{T}[\alpha \mid \langle n + m\rangle \setminus \alpha]$ add up to zero. That is, if we let

$$x = (x_1, x_2, \ldots x_n)^T, \quad x_i = \begin{cases} 1 & \text{if } i \in \alpha, \\ 0 & \text{otherwise,} \end{cases}$$

we have shown that

$$x^T[\hat{S}A\hat{S}\ \hat{S}B] = [x^T\ 0]$$

for an invertible signing $\hat{S}$. Hence, using $\lambda = 1$ in condition (2.3), it follows that $(\hat{S}A\hat{S}, \hat{S}B)$ and thus (A, B) is not sign controllable. ∎

COROLLARY 3.6. *Let $A \in \mathbf{R}^{n,\,n}$, $B \in \mathbf{R}^{n,\,m}$, $T = [A\ B]$, $\Gamma = \mathcal{D}(T)$, and suppose that the diagonal entries of A are nonzero and have the same sign. Let $\alpha \subseteq \langle n\rangle$ with $\alpha \subset \mathcal{R}(\alpha)$ such that $\mathcal{B}(T[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha])$ contains a nonsingular signing. Then (A, B) is not sign controllable.*

Proof. Let α be as prescribed and $S \in \mathcal{B}(T[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha])$ be nonsingular. Since all diagonal entries of $ST[\alpha]S$ are nonzero and have the same sign, all the assumptions of Proposition 3.5 are satisfied and the corollary follows. ∎

THEOREM 3.7. *Let $A \in \mathbf{R}^{n,\,n}$, $B \in \mathbf{R}^{n,\,m}$, $T = [A\ B]$, and $\Gamma = \mathcal{D}(T)$. Suppose that (A, B) is a strict linear control system. Then (A, B) is sign controllable if and only if the following conditions hold:*

(1) *Γ is accessible,*
(2) *$[A\ B]$ is an L-matrix, and*
(3) *for all $\alpha \subseteq \langle n\rangle$ satisfying $\alpha \subset \mathcal{R}(\alpha)$ in Γ, α is not aligned relative to $\mathcal{R}(\alpha) \setminus \alpha$.*

Proof. The sufficiency of conditions (1)–(3) follows from Theorem 3.2. We have discussed the necessity of conditions (1) and (2) after Lemma 3.1. To prove the necessity of condition (3), assume that (1) and (2) hold and that (3) does not hold.

Since Γ is accessible and the diagonal entries of A are nonzero, we have that for all $\alpha \subseteq \langle n \rangle$, $\alpha \subset \mathcal{R}(\alpha)$ and hence $\mathcal{R}(\alpha) \setminus \alpha \neq \emptyset$. So since (3) is not true, there exists $\alpha \subset \mathcal{R}(\alpha)$ such that $\mathcal{B}(T[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha]) \neq \emptyset$. Because (A, B) is strict, $\mathcal{B}(T[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha])$ must contain a nonsingular signing. By Corollary 3.6 it follows that (A, B) is not sign controllable. □

Example 3.8. Let

$$A = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

The directed graph of $T = [A\ B]$ is accessible and T is an L-matrix. For all $\alpha \subseteq \langle n \rangle$ we have that $\alpha \subset \mathcal{R}(\alpha)$ in Γ. In fact, for all $\alpha \subseteq \langle n \rangle$, except $\hat{\alpha} = \{1, 2, 3\}$, $T[\alpha \mid \mathcal{R}(\alpha) \setminus \alpha]$ is an L-matrix. We have that $T[\hat{\alpha} \mid \mathcal{R}(\hat{\alpha}) \setminus \hat{\alpha}] = B$, and $S = \operatorname{diag}(-1, 1, 1) \in \mathcal{B}(B)$. So (A, B) is a strict linear control system. Since $ST[\hat{\alpha}]$ has unisigned columns of only positive type, $\hat{\alpha}$ is aligned relative to $\mathcal{R}(\hat{\alpha}) \setminus \hat{\alpha}$. By Theorem 3.7, (A, B) is not sign controllable.

4. The extended controllability matrix. We will now introduce some additional concepts and terminology pertaining to an alternative analysis of sign controllability.

For the purposes of this section, we append to the set of signings the zero (square) matrix and refer to them as *weak signings*. We let $\mathcal{B}_0(X)$ denote $\mathcal{B}(X) \cup \{0\}$ for any $X \in \mathbf{R}^{s,\,t}$.

It is clear that for every $S \in \mathcal{B}_0(X)$ there exists $\hat{X} \in Q(X)$ such that the column sums of $S\hat{X}$ equal to zero (and hence equal to the column sums of the zero matrix). Based on this observation, we extend the notion of $\mathcal{B}_0(X)$ as follows. Given a matrix X and a weak signing S', we denote by $\mathcal{B}_0(X, S')$ the set of all weak signings S such that there exists $\hat{X} \in Q(X)$ with the column sums of $S\hat{X}$ equal to the column sums of S'. Notice that $\mathcal{B}_0(X) = \mathcal{B}_0(X, 0)$. To illustrate the definition of $\mathcal{B}_0(X, S')$, let

$$X = \begin{bmatrix} 1 & -3 & 2 \\ -1 & 0 & 1 \\ 2 & -1 & -1 \end{bmatrix} \quad \text{and} \quad S' = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then $\mathcal{B}_0(X, S')$ consists of all signings S such that SX has a negative entry in the first column, a positive entry in the second column, and a balanced third column. For example, $S = \operatorname{diag}(-1, 1, 1) \in \mathcal{B}_0(X, S')$ and $I \notin \mathcal{B}_0(X, S')$.

DEFINITION 4.1. *Let $A \in \mathbf{R}^{n,\,n}$ and $B \in \mathbf{R}^{n,\,m}$ be given. A nonzero ordered n-tuple $(S_1, S_2, \ldots, S_n)$ of weak signings is called an* (A, B)-balancing chain *if*

$$S_i \in \mathcal{B}_0(B) \quad (i = 1, 2, \ldots, n)$$

and

$$S_{i+1} \in \mathcal{B}_0(A, S_i) \quad (i = 1, 2, \ldots, n-1).$$

If, in addition, there exist $\hat{A} \in Q(A)$, $\hat{B} \in Q(B)$, and entrywise positive vectors $x_i \in \mathbf{R}^n$ such that

$$x_i^T S_i \hat{B} = 0 \quad (i = 1, 2, \ldots, n)$$

and

$$x_i^T S_i = x_{i+1}^T S_{i+1} \hat{A} \quad (i = 1, 2, \ldots, n-1),$$

we call $(S_1, S_2, \ldots, S_n)$ *a* compatible (A, B)*-balancing chain.*

The notion of an (A, B)-balancing chain depends only on the sign patterns of A and B. Indeed, if $(S_1, S_2, \ldots, S_n)$ is an (A, B)-balancing chain, then there always exist $A_i \in Q(A)$, $B_i \in Q(B)$, and x_i with positive entries such that $x_i^T S_i B_i = 0$ for $i = 1, 2, \ldots, n$ and $x_i^T S_i = x_{i+1}^T S_{i+1} A_{i+1}$ for $i = 1, 2, \ldots, n-1$. In fact, we can take $x_i = e$ for all i. In the definition of a compatible (A, B)-balancing chain we require, in addition, that there are common matrices $\hat{A} \in Q(A)$ and $\hat{B} \in Q(B)$ that satisfy the above conditions.

Observe that an (A, B)-balancing chain $(S_1, S_2, \ldots, S_n)$ may contain some zero weak signings, which could appear only as the leading part of the chain. Indeed, if $S_i \neq 0$, then since $S_{i+1}\hat{A}$ must have the same column sums as S_i for some $\hat{A} \in Q(A)$, S_{i+1} must be nonzero.

DEFINITION 4.2. *With the linear control system* (A, B) *we will associate (in Lemma* 4.3*) the* extended controllability matrix $\mathcal{G}$ *defined as follows:*

$$\mathcal{G} = \begin{bmatrix} I & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & B \\ -A & I & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & B & 0 \\ 0 & -A & I & \ldots & \ldots & \ldots & \ldots & 0 & B & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \iddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & -A & I & 0 & B & 0 & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & 0 & -A & B & 0 & \ldots & \ldots & \ldots & 0 \end{bmatrix} \in \mathbf{R}^{n^2,\ n(n+m-1)}.$$

The following result is a recasting of the classical condition for controllability in (2.2); its proof can be found in Casti [2].

LEMMA 4.3 (see [2, Corollary 5, section 3.5]). *Let* $A \in \mathbf{R}^{n,\ n}$, $B \in \mathbf{R}^{n,\ m}$. *The control system* (A, B) *is completely controllable if and only if rank* $\mathcal{G} = n^2$.

We now have the following equivalent condition for sign controllability.

THEOREM 4.4. *Let* $A \in \mathbf{R}^{n,\ n}$, $B \in \mathbf{R}^{n,\ m}$. *The linear control system* (A, B) *is sign controllable if and only if there is no compatible* (A, B)*-balancing chain.*

Proof. Let $A \in \mathbf{R}^{n,\ n}$, $B \in \mathbf{R}^{n,\ m}$, and suppose that (A, B) is not sign controllable. Then, by Lemma 4.3, there are matrices in $\hat{A} \in Q(A)$ and $\hat{B} \in Q(B)$ so that the corresponding extended controllability matrix $\hat{\mathcal{G}}$ is of deficient rank, i.e., $w^T \hat{\mathcal{G}} = 0$ for some $w \in \mathbf{R}^{n^2} \setminus \{0\}$. Now let $S \in \mathbf{R}^{n^2,\ n^2}$ be a signing such that $w = Sx$, where x has positive entries. It follows that $x^T S \hat{\mathcal{G}} = 0$. Hence if S is partitioned into n diagonal blocks $S_1, S_2, \ldots, S_n$ of size $n \times n$, then $(S_1, S_2, \ldots, S_n)$ is a compatible (A, B)-balancing chain.

Conversely, if $(S_1, S_2, \ldots, S_n)$ is a compatible (A, B)-balancing chain, then there exist $\hat{A} \in Q(A)$, $\hat{B} \in Q(B)$, and vectors x_i with positive entries such that

$$x_i^T S_i \hat{B} = 0 \quad (i = 1, 2, \ldots, n)$$

and

$$x_i^T S_i = x_{i+1}^T S_{i+1} \hat{A} \quad (i = 1, 2, \ldots, n-1).$$

It follows that for $S = \operatorname{diag}(S_1, S_2, \ldots, S_n)$ and for $x = [x_1^T, x_2^T, \ldots, x_n^T]^T$, $w = Sx$ is a nonzero left nullvector of the extended controllability matrix of $(\hat{A}, \hat{B})$; that is, by Lemma 4.3, (A, B) is not sign controllable. ∎

The existence or not of a compatible (A, B)-balancing chain can be a hard condition to check, but in some instances the clauses in the definition of a compatible (A, B)-balancing chain can serve as useful necessary or sufficient conditions. This is illustrated in the following examples.

Example 4.5. This example is mentioned in [4]. Let

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Notice that $\mathcal{B}_0(B)$ consists of the weak signings $S_1, S_2, \ldots, S_9$ having their (1,1) entry equal to zero. It is easy to check the sign patterns of $S_i A$ for $i = 1, 2, \ldots, 9$ and discover that there is no (A, B)-balancing chain and hence, by Theorem 4.4, (A, B) is sign controllable.

Example 4.6. Let

$$A = \begin{bmatrix} -1 & 0 & 1 \\ 1 & 0 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

It can be checked that

$$R = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad S = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}, \quad T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

are in $\mathcal{B}(B)$, and that (R, T, R), (T, R, T), (T, R, S) are some of the (A, B)-balancing chains. In this case the knowledge of a balancing chain leads to a straightforward search for the vectors x_i and the matrices $\hat{A}$ and $\hat{B}$ in the definition of a compatible balancing chain. One finds that with $x_1 = x_2 = x_3 = e$ and

$$\hat{A} = \begin{bmatrix} -1/3 & 0 & 1/3 \\ 1/3 & 0 & -1/3 \\ 1/3 & -1 & -1/3 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix},$$

(R, T, R) is a compatible (A, B)-balancing chain and thus, by Theorem 4.4, (A, B) is not sign controllable.

In conclusion, we have presented an alternative approach to sign controllability of a linear control system (A, B) based on the existence of a balancing chain of signings. We do not know if there exists an algorithm to verify the (non)existence of a compatible balancing chain, regardless of the complexity. We have also found sufficient conditions for sign controllability, based on the sign pattern and the directed graph of $[A\ B]$, which are necessary and sufficient when the linear control system is strict. We have not addressed computational matters regarding these conditions. However, we remark that the recognition of one of these conditions, namely, that the rectangular matrix $[A\ B]$ be an L-matrix, has been shown to be an NP-complete problem (see Klee, Ladner, and Manber [5]).

REFERENCES

[1] R. A. Brualdi and B. L. Shader, *Matrices of Sign-Solvable Linear Systems*, Cambridge University Press, London, 1995.

[2] J. L. Casti, *Dynamical Systems and Their Applications: Linear Theory*, Academic Press, New York, 1977.
[3] L. E. Faibusovich, *Algebraic Riccati equation and symplectic algebra*, Internat. J. Control, 43 (1986), pp. 781–792.
[4] C. R. Johnson, V. Mehrmann, and D. D. Olesky, *Sign controllability of a nonnegative matrix and a positive vector*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 398–407.
[5] V. Klee, R. Ladner, and R. Manber, *Signsolvability revisited*, Linear Algebra Appl., 59 (1984), pp. 131–157.
[6] C. T. Lin, *Structural Controllability*, IEEE Trans. Automat. Control, AC–19 (1974), pp. 201–208.
[7] P. Lancaster and L. Rodman, *Algebraic Riccati Equations*, Oxford University Press, New York, 1995.
[8] K. Murota, *Systems Analysis by Graphs and Matroids–Structural Solvability and Controllability*, Algorithms and Combinatorics 3, Springer-Verlag, Berlin, Heidelberg, 1987.
[9] H. Mayeda and T. Yamada, *Strong structural controllability*, SIAM J. Control Optim., 17 (1979), pp. 123–138.
[10] D. D. Olesky, M. Tsatsomeros, and P. van den Driessche, *Qualitative controllability and uncontrollability by a single entry*, Linear Algebra Appl., 187 (1993), pp. 183–194.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 365–368, April 1998

005

MORE ON CONCAVITY OF A MATRIX FUNCTION*

JÜRGEN GROSS†

Abstract. The mapping $\mathbf{A} \mapsto (\mathbf{K}'\mathbf{A}^+\mathbf{K})^+$ is shown to be matrix concave and isoton when $\mathbf{A}$ varies over the set of symmetric nonnegative definite matrices whose range is invariant with respect to $\mathbf{K}\mathbf{K}'$. This generalizes a well-known result in statistical literature.

Key words. matrix-concave function, Löwner partial ordering, linear model

AMS subject classifications. 15A09, 15A45, 15A03, 62J12

PII. S0895479896311244

1. Introduction. Let $\mathbb{R}^{m\times n}$ denote the set of $m \times n$ real matrices. The symbols $\mathbf{A}'$, $\mathbf{A}^-$, $\mathbf{A}^+$, $\mathrm{rk}(\mathbf{A})$, $\mathcal{R}(\mathbf{A})$, and $\mathcal{N}(\mathbf{A})$ will stand for the transpose, any generalized inverse, the Moore–Penrose inverse, the rank, the range, and the nullspace, respectively, of $\mathbf{A} \in \mathbb{R}^{m\times n}$. By $\mathbf{A}^\perp$ we denote any matrix whose range coincides with the orthogonal complement of $\mathcal{R}(\mathbf{A})$. A possible choice for $\mathbf{A}^\perp$ is $\mathbf{A}^\perp = \mathbf{I}_m - \mathbf{A}\mathbf{A}^+$. For any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times m}$ we will write $\mathbf{A} \overset{\text{L}}{\leq} \mathbf{B}$ when $\mathbf{B} - \mathbf{A} = \mathbf{G}\mathbf{G}'$ for some matrix $\mathbf{G}$, or, in other words, $\mathbf{B} - \mathbf{A}$ is symmetric nonnegative definite. Note that according to Löwner (1934), the relation $\overset{\text{L}}{\leq}$ specifies a partial ordering in $\mathbb{R}^{m\times m}$. For any matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ with real eigenvalues, $\lambda_{\max}(\mathbf{A})$ denotes the maximal eigenvalue of $\mathbf{A}$. Observe that all eigenvalues of a symmetric matrix are real.

In a recent note, Neudecker and Liu (1995) give an algebraic proof for the concavity of the matrix function $f(\mathbf{A}) = (\mathbf{K}'\mathbf{A}^-\mathbf{K})^+$, where $\mathbf{A}$ belongs to the convex cone $\mathcal{A}(\mathbf{K}) = \{\mathbf{0} \overset{\text{L}}{\leq} \mathbf{A} : \mathcal{R}(\mathbf{K}) \subseteq \mathcal{R}(\mathbf{A})\}$ and $\mathbf{K} \in \mathbb{R}^{m\times p}$. This has originally been established by Pukelsheim and Styan (1983), who apply their result to a combination of estimators from two linear models with different dispersion matrices. Another proof is given by Gaffke and Krafft (1982, Theorem 4.8), who refer the result to an earlier version of Pukelsheim and Styan's (1983) paper.

However, from an algebraical point of view, one may also consider a more general function for which analogous properties can be derived. As we will show subsequently, application to linear models is possible.

In the following we will consider the matrix function $f^*(\mathbf{A}) = (\mathbf{K}'\mathbf{A}^+\mathbf{K})^+$, where $\mathbf{A}$ belongs to

$$\mathcal{A}^*(\mathbf{K}) = \{\mathbf{0} \overset{\text{L}}{\leq} \mathbf{A} : \mathcal{R}(\mathbf{K}\mathbf{K}'\mathbf{A}) \subseteq \mathcal{R}(\mathbf{A})\}.$$

Since for $\mathbf{0} \overset{\text{L}}{\leq} \mathbf{A}$ and $\mathbf{0} \overset{\text{L}}{\leq} \mathbf{B}$ we have $\mathcal{R}(\mathbf{A}) + \mathcal{R}(\mathbf{B}) = \mathcal{R}(\mathbf{A} + \mathbf{B})$, it is easy to see that $\mathcal{A}^*(\mathbf{K})$ is a convex cone, i.e., $\alpha\mathbf{A}$ and $\mathbf{A} + \mathbf{B}$ lie in $\mathcal{A}^*(\mathbf{K})$ for all $\mathbf{A}, \mathbf{B} \in \mathcal{A}^*(\mathbf{K})$, $\alpha > 0$.

The difference between $f(\mathbf{A})$ and $f^*(\mathbf{A})$ lies in the fact that

$$f(\mathbf{A}) = (\mathbf{K}'\mathbf{A}^-\mathbf{K})^+ = (\mathbf{K}'\mathbf{A}^+\mathbf{K})^+$$

*Received by the editors October 23, 1996; accepted for publication (in revised form) by G. P. Styan March 5, 1997. This research was supported by Deutsche Forschungsgemeinschaft grant Tr 253/2-1/2-2.

http://www.siam.org/journals/simax/19-2/31124.html

†Department of Statistics, University of Dortmund, Vogelpothsweg 87, D-44221 Dortmund, Germany (gross@amadeus.statistik.uni-dortmund.de).

for all generalized inverses $\mathbf{A}^-$ of $\mathbf{A}$ whenever $\mathbf{A} \in \mathcal{A}(\mathbf{K})$, whereas $\mathbf{A}^+$ in $f^*(\mathbf{A}) = (\mathbf{K}'\mathbf{A}^+\mathbf{K})^+$ cannot be replaced by $\mathbf{A}^-$ when $\mathbf{A} \in \mathcal{A}^*(\mathbf{K})$. However, it is clear that $\mathcal{A}(\mathbf{K}) \subseteq \mathcal{A}^*(\mathbf{K})$ for a fixed matrix $\mathbf{K}$.

On the other hand, both functions can be seen as generalizations of the function $\phi(\mathbf{A}) = (\mathbf{K}'\mathbf{A}^{-1}\mathbf{K})^{-1}$, where $\mathbf{A}$ is assumed to be nonsingular and $\mathbf{K}$ has full column rank. The function ϕ has been considered by Marshall and Olkin (1979, pp. 469–473).

In the next section we demonstrate that the positively homogeneous function $f^*(\mathbf{A})$ is matrix concave with respect to the Löwner ordering by showing that

$$f^*(\mathbf{A}) + f^*(\mathbf{B}) \overset{\text{L}}{\leq} f^*(\mathbf{A} + \mathbf{B})$$

for all matrices $\mathbf{A}$ and $\mathbf{B}$ in the convex cone $\mathcal{A}^*(\mathbf{K})$.

2. Results. The following characterization of the Löwner partial ordering of symmetric nonnegative definite matrices, originally established by Stępniak (1985), is quite useful. For a tractable proof see Liski and Puntanen (1989, Lemma). A more general statement is given by Baksalary, Schipp, and Trenkler (1992, Theorem 1).

LEMMA 1. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times m}$ *be such that* $\mathbf{0} \overset{L}{\leq} \mathbf{A}$ *and* $\mathbf{0} \overset{L}{\leq} \mathbf{B}$. *Then* $\mathbf{A} \overset{L}{\leq} \mathbf{B}$ *if and only if* $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$ *and* $\lambda_{\max}(\mathbf{A}\mathbf{B}^-) \leq 1$.

Note that in the above lemma, $\lambda_{\max}(\mathbf{A}\mathbf{B}^-)$ is invariant with respect to the choice of $\mathbf{B}^-$, i.e., $\lambda_{\max}(\mathbf{A}\mathbf{B}^-) = \lambda_{\max}(\mathbf{A}\mathbf{B}^+)$ for all generalized inverses $\mathbf{B}^-$ of $\mathbf{B}$. This is easily established by using $\mathbf{B}\mathbf{B}^-\mathbf{A} = \mathbf{A}$, and the fact that the set of nontrivial eigenvalues of a matrix product $\mathbf{F}\mathbf{G}$ coincides with the set of nontrivial eigenvalues of $\mathbf{G}\mathbf{F}$.

In our Theorem 1 we will show matrix concavity of f^* with the help of Lemma 1. Before stating this result, we give two further lemmas which provide some insight concerning the generalization of the function f to the function f^* on the convex cones $\mathcal{A}(\mathbf{K})$ and $\mathcal{A}^*(\mathbf{K})$, respectively.

LEMMA 2. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times m}$ *be such that* $\mathbf{0} \overset{L}{\leq} \mathbf{A}$ *and* $\mathbf{0} \overset{L}{\leq} \mathbf{B}$ *and let* $\mathbf{K} \in \mathbb{R}^{m\times p}$. *Then*

$$\mathcal{R}[(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+ + (\mathbf{K}'\mathbf{B}^+\mathbf{K})^+] = \mathcal{R}[(\mathbf{K}'(\mathbf{A}+\mathbf{B})^+\mathbf{K})^+].$$

Proof. Since $(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+$ and $(\mathbf{K}'\mathbf{B}^+\mathbf{K})^+$ are symmetric nonnegative definite, we have $\mathcal{R}[(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+ + (\mathbf{K}'\mathbf{B}^+\mathbf{K})^+] = \mathcal{R}[(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+] + \mathcal{R}[(\mathbf{K}'\mathbf{B}^+\mathbf{K})^+] = \mathcal{R}(\mathbf{K}'\mathbf{A}) + \mathcal{R}(\mathbf{K}'\mathbf{B}) = \mathcal{R}[(\mathbf{K}'(\mathbf{A}+\mathbf{B})] = \mathcal{R}[(\mathbf{K}'(\mathbf{A}+\mathbf{B})^+\mathbf{K})^+]$. □

Clearly Lemma 2 implies $\mathcal{R}[f^*(\mathbf{A}) + f^*(\mathbf{B})] = \mathcal{R}[f^*(\mathbf{A}+\mathbf{B})]$ for $\mathbf{A}, \mathbf{B} \in \mathcal{A}^*(\mathbf{K})$.

LEMMA 3. *Let* $\mathbf{A} \in \mathbb{R}^{m\times m}$ *be such that* $\mathbf{0} \overset{L}{\leq} \mathbf{A}$ *and let* $\mathbf{K} \in \mathbb{R}^{m\times p}$. *Then* $\mathbf{K}(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}' \overset{L}{\leq} \mathbf{A}$ *if and only if* $\mathcal{R}(\mathbf{K}\mathbf{K}'\mathbf{A}) \subseteq \mathcal{R}(\mathbf{A})$.

Proof. Since the matrix $\mathbf{K}(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}'\mathbf{A}^+$ is idempotent, it has only eigenvalues 0 and 1, which shows $\lambda_{\max}[\mathbf{K}(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}'\mathbf{A}^+] \overset{\text{L}}{\leq} 1$. Moreover, $\mathcal{R}[\mathbf{K}(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}'] = \mathcal{R}(\mathbf{K}\mathbf{K}'\mathbf{A})$ and the assertion follows from Lemma 1. □

Using Lemmas 1, 2, and 3, the following theorem is easy to establish.

THEOREM 1. *Let* $\mathbf{A}, \mathbf{B} \in \mathcal{A}^*(\mathbf{K})$. *Then* $f^*(\mathbf{A}) + f^*(\mathbf{B}) \overset{L}{\leq} f^*(\mathbf{A}+\mathbf{B})$.

Proof. From Lemma 3 we immediately get

$$\mathbf{K}(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}' + \mathbf{K}(\mathbf{K}'\mathbf{B}^+\mathbf{K})^+\mathbf{K}' \overset{\text{L}}{\leq} \mathbf{A}+\mathbf{B}.$$

Then from Lemma 1

$$\lambda_{\max}\{[\mathbf{K}(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}' + \mathbf{K}(\mathbf{K}'\mathbf{B}^+\mathbf{K})^+\mathbf{K}'](\mathbf{A}+\mathbf{B})^+\} \leq 1,$$

which is equivalent to $\lambda_{\max}\{[f^*(\mathbf{A})+f^*(\mathbf{B})][f^*(\mathbf{A}+\mathbf{B})]^+\} \le 1$. Applying Lemmas 1 and 2 gives the assertion. □

The following result establishes the fact that on the convex cone $\mathcal{A}^*(\mathbf{K})$ the function f^* is not only concave but also isoton with respect to the Löwner partial ordering. Compare also Theorem 3 in Pukelsheim and Styan (1983).

THEOREM 2. *Let* $\mathbf{A},\mathbf{B} \in \mathcal{A}^*(\mathbf{K})$ *be such that* $\mathbf{A} \overset{L}{\le} \mathbf{B}$. *Then* $f^*(\mathbf{A}) \overset{L}{\le} f^*(\mathbf{B})$.

Proof. When $\mathbf{A} \overset{L}{\le} \mathbf{B}$, Lemma 1 gives $\mathcal{R}(\mathbf{A}) \subseteq \mathcal{R}(\mathbf{B})$, from which follows $\mathcal{R}(\mathbf{K}'\mathbf{A}) \subseteq \mathcal{R}(\mathbf{K}'\mathbf{B})$, i.e., $\mathcal{R}[f^*(\mathbf{A})] \subseteq \mathcal{R}[f^*(\mathbf{B})]$. Moreover, from Lemma 3 we get $\mathbf{K}(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}' \overset{L}{\le} \mathbf{B}$, which implies $\lambda_{\max}[\mathbf{K}(\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}'\mathbf{B}^+] \le 1$ in view of Lemma 1. Since the latter is equivalent to $\lambda_{\max}\{f^*(\mathbf{A})[f^*(\mathbf{B})]^+\} \le 1$, the assertion follows with Lemma 1. □

Observe that according to Theorem 3 in Pukelheim and Styan (1983) (see also Proposition 2 in Neudecker and Liu (1995)), $f(\mathbf{A}) \overset{L}{\le} f(\mathbf{B})$ is equivalent to $[f(\mathbf{B})]^+ \overset{L}{\le} [f(\mathbf{A})]^+$, i.e., Moore–Penrose inversion of the function f is antiton with respect to the Löwner partial ordering. It should be pointed out that an analogous statement does not hold for the function f^*. From Milliken and Akdeniz (1977, Theorem 3.1) we know that $f^*(\mathbf{A}) \overset{L}{\le} f^*(\mathbf{B})$ and $[f^*(\mathbf{B})]^+ \overset{L}{\le} [f^*(\mathbf{A})]^+$ hold together if and only if $\mathrm{rk}[f^*(\mathbf{A})] = \mathrm{rk}[f^*(\mathbf{B})]$, i.e., $\mathrm{rk}(\mathbf{A}\mathbf{K}) = \mathrm{rk}(\mathbf{B}\mathbf{K})$. However, the possible choice $\mathbf{K} = \mathbf{I}_m$, $p = m$, shows that the latter is not necessarily satisfied, even under the assumptions of Theorem 2.

According to Hartwig [1978, Theorem 1i], who established his result independently of Milliken and Akdeniz (1977), the condition $\mathrm{rk}[f^*(\mathbf{A})] = \mathrm{rk}[f^*(\mathbf{B})]$ can be replaced by $\mathcal{R}[f^*(\mathbf{A})] = \mathcal{R}[f^*(\mathbf{B})]$, i.e., $\mathcal{R}(\mathbf{K}'\mathbf{A}) = \mathcal{R}(\mathbf{K}'\mathbf{B})$. The following theorem is concerned with giving an equivalent condition, which depends on the subspaces $\mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{B})$ themselves.

THEOREM 3. *Let* $\mathbf{A},\mathbf{B} \in \mathcal{A}^*(\mathbf{K})$ *be such that* $f^*(\mathbf{A}) \overset{L}{\le} f^*(\mathbf{B})$. *Then* $[f^*(\mathbf{B})]^+ \overset{L}{\le} [f^*(\mathbf{A})]^+$ *if and only if*

$$\mathcal{R}(\mathbf{A}) \cap \mathcal{R}(\mathbf{K}) = \mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{K}).$$

Proof. By applying Theorem 1i in Hartwig (1978) we get $[f^*(\mathbf{B})]^+ \overset{L}{\le} [f^*(\mathbf{A})]^+$ if and only if $\mathcal{R}[f^*(\mathbf{A})] = \mathcal{R}[f^*(\mathbf{B})]$, i.e., $\mathcal{R}(\mathbf{K}'\mathbf{A}) = \mathcal{R}(\mathbf{K}'\mathbf{B})$, i.e., $\mathcal{R}(\mathbf{K}\mathbf{K}'\mathbf{A}) = \mathcal{R}(\mathbf{K}\mathbf{K}'\mathbf{B})$. But since $\mathbf{A} \in \mathcal{A}^*(\mathbf{K})$, $\mathcal{R}(\mathbf{K}\mathbf{K}'\mathbf{A}) \subseteq \mathcal{R}(\mathbf{A}) \cap \mathcal{R}(\mathbf{K})$. Now, from the strengthened version of Sylvester's law of nullity, cf. Baksalary and Styan (1993, Corollary 1), we always have $\mathrm{rk}(\mathbf{K}\mathbf{K}'\mathbf{A}) \ge \dim[\mathcal{R}(\mathbf{A}) \cap \mathcal{R}(\mathbf{K})]$, showing that $\mathcal{R}(\mathbf{K}\mathbf{K}'\mathbf{A}) = \mathcal{R}(\mathbf{A}) \cap \mathcal{R}(\mathbf{K})$. Analogously, $\mathcal{R}(\mathbf{K}\mathbf{K}'\mathbf{B}) = \mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{K})$ and the assertion follows. □

Clearly when $\mathcal{R}(\mathbf{K}) \subseteq \mathcal{R}(\mathbf{A})$ and $\mathcal{R}(\mathbf{K}) \subseteq \mathcal{R}(\mathbf{B})$, the condition $\mathcal{R}(\mathbf{A}) \cap \mathcal{R}(\mathbf{K}) = \mathcal{R}(\mathbf{B}) \cap \mathcal{R}(\mathbf{K})$ reduces to $\mathcal{R}(\mathbf{K}) = \mathcal{R}(\mathbf{K})$. This covers the above-mentioned results of Pukelsheim and Styan (1983, Theorem 3) and Neudecker and Liu (1995, Proposition 2).

Note that the equivalence of the conditions $\mathcal{R}(\mathbf{F}\mathbf{G}) \subseteq \mathcal{R}(\mathbf{G})$ and $\mathcal{R}(\mathbf{F}\mathbf{G}) = \mathcal{R}(\mathbf{G}) \cap \mathcal{R}(\mathbf{F})$ for symmetric nonnegative definite matrix $\mathbf{F}$ and arbitrary matrix $\mathbf{G}$ also appears in connection with equality conditions for ordinary least squares and best linear unbiased estimators, cf. Alalouf and Styan [1984, Theorem 2, equations (2.5), (2.6)].

3. Application. Consider now the linear model $\{\mathbf{y}, \mathbf{K}\beta, \mathbf{A}\}$, where $\mathbf{y}$ is an $m \times 1$ random vector with mean vector $\mathbf{K}\beta$ and dispersion matrix $\mathbf{A} \in \mathcal{A}^*(\mathbf{K})$. Then

Theorem 4 in Baksalary (1987) asserts that $\mathbf{KHy}$ with

$$\mathbf{H} = (\mathbf{K}'\mathbf{A}^+\mathbf{K})^+\mathbf{K}'\mathbf{A}^+ + (\mathbf{K}'(\mathbf{I}_m - \mathbf{A}\mathbf{A}^+)\mathbf{K})^+\mathbf{K}'(\mathbf{I}_m - \mathbf{A}\mathbf{A}^+)$$

is the best linear unbiased estimator for $\mathbf{K}\beta$, which means $\mathbf{KHK} = \mathbf{K}$ and $\mathbf{KHAK}^\perp = \mathbf{0}$, cf. Rao [1978, (3.1)]. But since $\mathcal{R}(\mathbf{H}) \subseteq \mathcal{R}(\mathbf{K}')$ this also gives $\mathbf{HKK}' = \mathbf{K}'$ and $\mathbf{HAK}^\perp = \mathbf{0}$, showing that $\mathbf{Hy}$ is best linear minimum biased for β, cf. Rao [1978, (3.21)]. Clearly $\mathbf{Hy}$ has dispersion matrix $f^*(\mathbf{A})$.

Consider in addition a linear model $\{\mathbf{z}, \mathbf{K}\beta, \mathbf{B}\}$, $\mathbf{B} \in \mathcal{A}^*(\mathbf{K})$, where $\mathbf{z}$ and $\mathbf{y}$ are uncorrelated. By applying our Theorem 1, the same reasoning as in Pukelsheim and Styan (1983, section 3) shows that averaging the individual best linear minimum biased estimators for β is preferable to averaging the observations in advance and estimating afterwards.

REFERENCES

I.S. Alalouf and G.P.H. Styan (1984), *Characterizations of the conditions for the ordinary least squares estimator to be best linear unbiased*, in Topics in Applied Statistics, Y.P. Chaubey and T.D. Dwivedi, eds., Concordia University, Department of Mathematics, Montréal, pp. 331–344.

J.K. Baksalary (1987), *Algebraic characterizations and statistical implications of the commutativity of orthogonal projectors*, in Proc. Second International Tampere Conference in Statistics, T. Pukkila and S. Puntanen, eds., University of Tampere, Tampere, pp. 113–142.

J.K. Baksalary, B. Schipp, and G. Trenkler (1992), *Some further results on hermitian-matrix inequalities*, Linear Algebra Appl., 160, pp. 119–129.

J.K. Baksalary and G.P.H. Styan (1993), *Around a formula for the rank of a matrix product with some statistical applications*, in Graphs, Matrices, and Designs, Festschrift in Honor of Norman J. Pullman, R.S. Rees, eds., Marcel Dekker, New York, pp. 1–18.

N. Gaffke and O. Krafft (1982), *Matrix inequalities in the Löwner ordering*, in Modern Applied Mathematics, B. Korte, eds., North–Holland, Amsterdam, pp. 595–622.

R.E. Hartwig (1978), *A note on the partial ordering of positive semi-definite matrices*, Linear and Multilinear Algebra, 6, pp. 223–226.

E.P. Liski and S. Puntanen (1989), *A further note on a theorem on the difference of the generalized inverses of two nonnegative definite matrices*, Comm. Statist. Theory Methods, 18, pp. 1747–1751.

K. Löwner (1934), *Über monotone Matrixfunktionen*, Math. Z., 38, pp. 177–216.

A.W. Marshall and I. Olkin (1979), *Inequalities: Theory of Majorization and Its Applications*. Academic Press, San Diego, CA.

G.A. Milliken and F. Akdeniz (1977), *A theorem on the difference of the generalized inverses of two nonnegative matrices*, Comm. Statist. Theory Methods, 6, pp. 73–79.

H. Neudecker and S. Liu (1995), *Note on a matrix-concave function*, J. Math. Anal. Appl., 196, pp. 1139–1141.

F. Pukelsheim and G.P.H. Styan (1983), *Convexity and monotonicity properties of dispersion matrices of estimators in linear models*, Scand. J. Statist., 10, pp. 145–149.

C.R. Rao (1978), *Choice of best linear estimators in the Gauss-Markoff model with a singular dispersion matrix*, Comm. Statist. Theory Methods, 7, pp. 1199–1208.

C. Stępniak (1985), *Ordering of nonnegative definite matrices with application to comparison of linear models*, Linear Algebra Appl., 70, pp. 67–71.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 369–377, April 1998

006

ON THE RELATIONSHIP BETWEEN GRAPHS AND TOTALLY POSITIVE MATRICES*

J. M. PEÑA†

Abstract. A real matrix is said to be *totally positive* if all its minors are nonnegative. In this paper it is shown that properties of totally positive matrices can be applied to graph theory, and conversely. In fact, some properties of undirected and directed graphs (digraphs) are characterized in terms of the associated totally positive matrices. Some results on the existence of nonintersecting paths in a digraph are also provided.

Key words. totally positive matrix, graph, digraph, nonintersecting paths

AMS subject classifications. 15A48, 05C20, 05C50

PII. S089547989630396X

1. Introduction.

The relations between graph theory and matrix theory constitute a well-established area of research (see [7]). This paper explores some connections between theoretic properties of totally positive matrices and graph theoretic properties of certain graphs naturally associated with the matrices.

Section 2 deals with undirected graphs. Given a symmetric matrix $A = (a_{ij})_{1\leq i,j\leq n}$, the undirected graph $G(A)$ is the usual graph in which there is an edge $\{i, j\}$ if and only if $i \neq j$ and $a_{ij} \neq 0$. We call a *clique* of a graph a vertex-induced subgraph of G that is complete (i.e., all possible pairs of different vertices are edges). The maximum cardinality of a clique in G will be denoted by $c(G)$. If $c(G) = 2$, we shall say that the graph is *triangle-free*. There are many examples of triangle-free graphs. Obviously, trees and cycles are triangle-free. Since a cycle of a bipartite graph has even length, bipartite graphs are also triangle-free. A graph without simple cycles of length greater than or equal to four is usually said to be *chordal*. We say that a graph is *quadrilateral-free* if it has no cycles of length four.

Let us introduce now some of the classes of matrices that will be used in this paper. An $n \times n$ matrix A is TP_k if all $r \times r$ minors of A are nonnegative for all $r = 1, \ldots, k$. If A is TP_n, then it is called *totally positive*. This class of matrices has many applications in mathematics, statistics, economics, etc. (see [14], [1]). Some recent characterizations of totally positive matrices can be found in [9], [10], [11]. In Proposition 2.1 we prove that a symmetric TP_2 matrix A with nonzero rows is p-banded if and only if $c(G(A)) \leq p$; so, in particular, $G(A)$ is triangle-free if and only if A is tridiagonal. The corresponding characterization of the symmetric TP_2 matrices A such that $G(A)$ is quadrilateral-free is given in Proposition 2.4.

A square matrix A is called *M-matrix* if $A = \alpha I - P$, where I is the identity matrix, P is a nonnegative matrix, and $\alpha \geq \rho(P)$ (the spectral radius of P) is a positive real number. M-matrices have many equivalent definitions (see [5, Chapter 6]). M-matrices have important applications, for instance, in iterative methods in numerical analysis, in the analysis of dynamical systems, in economics, and in mathematical

*Received by the editors May 20, 1996; accepted for publication (in revised form) by T. Ando March 10, 1997.

http://www.siam.org/journals/simax/19-2/30396.html

†Departamento de Matemática Aplicada, Universidad de Zaragoza, 50009 Zaragoza, Spain (jmpena@posta.unizar.es). The work of this author was partially supported by DGICYT PB93-0310.

programming. Given a matrix $A = (a_{ij})_{1\leq i,j\leq n}$, its *comparison matrix* is defined by $\mathcal{M}(A) = (m_{ij})_{1\leq i,j\leq n}$ (with $m_{ii} := |a_{ii}|$ and $m_{ij} := -|a_{ij}|$ if $i \neq j$, $1 \leq i, j \leq n$).

An $n \times n$ matrix A is *completely positive* if it can be written as $A = BB^T$, where B is an $n \times m$ nonnegative matrix. Completely positive matrices are positive semidefinite matrices and, for any graph G, there exists a completely positive matrix A such that $G(A) = G$. Some recent results on graphs which are associated with completely positive matrices can be found in [3] and [4], and [2] surveys many results on completely positive matrices. [8, Theorem 5] characterizes completely positive matrices A with $G(A)$ triangle-free and [13, Theorem 1] characterizes M-matrices A with $G(A)$ triangle-free. In Proposition 2.5 we characterize the corresponding case of totally positive matrices. At the end of section 2 some results on the convergence of iterative methods to solve tridiagonal totally positive linear systems are given.

Section 3 deals with directed graphs (digraphs). The existence of nonintersecting paths in a digraph is a topic of wide interest in combinatorics. We give several results on this topic. Our fundamental tools to obtain these results are provided by an interpretation of the totally positive matrices in terms of digraphs (which was given in [6, Theorem 3.1]) and some properties of the totally positive matrices obtained in [9], [10], and [12].

2. Undirected graphs and totally positive matrices. In this section we shall deal with undirected graphs on vertices $\{1, 2, \ldots, n\}$. Now let us introduce some matricial notation. An $n \times n$ matrix is *p-banded* if all its entries are zero except within the band $|i - j| < p$. A tridiagonal matrix is a 2-banded matrix. Given $k, n \in \mathbf{N}$, $k \leq n$, we define $Q_{k,n} := \{(\alpha_1, \ldots, \alpha_k) | \alpha_i \in \mathbf{N}, 1 \leq \alpha_1 < \cdots < \alpha_k \leq n\}$ and for $\alpha, \beta \in Q_{k,n}$, $A[\alpha|\beta]$ is by definition the $k \times k$ submatrix of A containing rows numbered by α and columns numbered by β. Finally, $A[\alpha] := A[\alpha|\alpha]$.

In the next result we shall use the condition that the matrix has no zero rows because we are interested in graphs with no isolated vertices.

PROPOSITION 2.1. *Let A be a symmetric TP_2 matrix with nonzero rows. Then $c(G(A)) \leq p$ if and only if A is p-banded.*

Proof. As every $n \times n$ matrix is n-banded, we assume $p \leq n - 1$. If A is p-banded, then $c(G(A)) \leq p$ since A has no $(p + 1) \times (p + 1)$ principal submatrices whose off-diagonal entries are nonzero.

Let us assume now that A is an $n \times n$ matrix with $c(G(A)) \leq p$ and let us see by induction on n that A is p-banded. If $n = 2$ and $c(G(A)) = 1$, then A has a zero off-diagonal entry and, by symmetry, A is a diagonal matrix. Let us suppose that the result holds for $n - 1$ and let us prove it for n. By the induction hypothesis, the matrices $A[1, \ldots, n-1]$ and $A[2, \ldots, n]$ are already p-banded. Thus, it remains to see $a_{1n} = 0$ (and so, by symmetry, $a_{n1} = 0$). Let us assume that $a_{1n} \neq 0$ and we shall get a contradiction.

If $p < n - 1$, we have that $a_{1,n-1} = 0$. Since by symmetry A has no zero columns, there exists $k \in \{2, \ldots, n\}$ such that $a_{k,n-1} > 0$. Thus, $\det A[1, k|n-1, n] = -a_{1n}a_{k,n-1} < 0$, which contradicts that A is TP_2.

Finally, let us consider the case of $p = n - 1$. Let us observe that arguments similar to those of the previous case show that $a_{1k} \neq 0$ for all k and that $a_{kn} \neq 0$ for all k. Since $c(G(A)) \leq p$, $p = n - 1$, and A is symmetric, there exists $a_{ij} = 0$ for some $i \leq j < n$. So then $\det A[1, i|1, j] = -a_{i1}a_{1j} < 0$, which gives us the final contradiction. □

Applying the previous result to $p = 2$, we derive the following characterization of symmetric TP_2 matrices A with $G(A)$ triangle-free.

COROLLARY 2.2. *Let A be a symmetric TP_2 matrix with nonzero rows. Then $G(A)$ is triangle-free if and only if A is tridiagonal.*

We have already mentioned that the condition of dealing with matrices without zero rows is natural in the framework of graphs. Let us observe that without this restriction the previous result does not hold because, for instance, the symmetric matrix

$$A = \begin{pmatrix} 3 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 3 \end{pmatrix}$$

is totally positive and $G(A)$ is triangle-free, but A is not tridiagonal.

Now we can easily deduce from Corollary 2.2 a characterization of the TP_2 symmetric matrices A such that $G(A)$ is a tree.

COROLLARY 2.3. *Let A be a symmetric TP_2 matrix. Then the following conditions are equivalent:*

(i) *$G(A)$ is a tree.*

(ii) *A is tridiagonal and irreducible.*

(iii) *$G(A)$ is triangle-free and connected.*

The next result characterizes the case when $G(A)$ is quadrilateral-free.

PROPOSITION 2.4. *Let $A = (a_{ij})_{1\le i,j\le n}$ be a symmetric TP_2 matrix with nonzero rows. Then $G(A)$ is quadrilateral-free if and only if A is 3-banded and for every i $(1 \le i \le n-3)$ such that $a_{i,i+2} \neq 0$ one has that $a_{i+1,i+3} = 0$.*

Proof. Let us assume first that $G(A)$ is quadrilateral-free. Since $c(G(A)) \le 3$, A is 3-banded by Proposition 2.1. Let us suppose now that there exists an index i such that $a_{i,i+2} \neq 0$ and $a_{i+1,i+3} \neq 0$ and we shall obtain a contradiction. By symmetry, $a_{i+2,i} \neq 0$ and $a_{i+3,i+1} \neq 0$, and these four numbers are positive since A is TP_2. If $a_{i,i+1} = 0$, we would have $\det A[i, i+3|i+1, i+2] = -a_{i,i+2}a_{i+3,i+1} < 0$, which contradicts that A is TP_2. Analogously, if $a_{i+2,i+3} = 0$, we would have $\det A[i+1, i+2|i, i+3] = -a_{i+1,i+3}a_{i+2,i} < 0$, which contradicts again that A is TP_2. In conclusion, the elements $a_{i,i+2}, a_{i+1,i+3}, a_{i,i+1}$, and $a_{i+2,i+3}$ are nonzero and so they are associated with a cycle of length four in $G(A)$, which again gives a contradiction.

Let us prove now the converse. Let us assume that A is 3-banded and that $G(A)$ has a cycle of length four. Let $i(\le n-3)$ be the least index associated with the vertices of this cycle. Since A is 3-banded, the other two vertices of the cycle adjacent to i must be associated with the indices $i+1$ and $i+2$, and $i+3$ must correspond to the fourth vertex, which is adjacent to $i+1$ and $i+2$. But then we have that $a_{i,i+2} \neq 0$ and $a_{i+1,i+3} \neq 0$, which proves the converse of the proposition. ∎

In [8, Theorem 5] completely positive matrices A with $G(A)$ triangle-free were characterized, and a characterization of nonsingular M-matrices with $G(A)$ triangle-free was obtained in [13, Theorem 1]. The next result gives the corresponding characterization for the case of symmetric totally positive matrices.

PROPOSITION 2.5. *Let A be a nonsingular, nonnegative, and symmetric matrix with $G(A)$ triangle-free. Then the following conditions are equivalent:*

(i) *A is totally positive.*

(ii) *A is tridiagonal completely positive.*

(iii) *$\mathcal{M}(A)$ is a tridiagonal M-matrix.*

Proof. (i) $\Longrightarrow$ (ii) If A is totally positive and $G(A)$ is triangle-free, then A is tridiagonal by Corollary 2.2. On the other hand, a totally positive matrix admits an LU factorization with L and U totally positive (cf. [1, Theorem 3.5]) and since A is symmetric we can deduce that it is completely positive. In fact, the factors in the

unique LDU factorization of A are nonnegative (see, for instance, the second part of Theorem 4.1′ of [11]) and therefore the factors in the Cholesky factorization of A are nonnegative matrices.

(ii) $\Longrightarrow$ (i) If A is completely positive, it is in particular a positive semidefinite symmetric matrix and so its principal minors are nonnegative. Now (i) follows from [1, Theorem 2.3] since A is a nonnegative tridiagonal matrix.

(ii) $\Longleftrightarrow$ (iii) It is a consequence of [8, Theorem 5]. □

REMARK 2.6. *In contrast with the previous results, condition* (i) *of Proposition* 2.5 *imposes the total positivity of the matrix A instead of the property of being TP_2. The following matrix A shows that a TP_2 tridiagonal matrix is not necessarily totally positive:*

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 2 \\ 0 & 2 & 4 \end{pmatrix}$$

($\det A < 0$). On the other hand, Proposition 2.5 *cannot be extended in a natural way to matrices A with $G(A)$ quadrilateral-free. For instance, the matrix*

$$B = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix} \left(= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right)$$

is a completely positive 3-banded symmetric matrix with nonzero rows, but B is not totally positive (it is not even TP_2).

In the next result we see that the equivalence between (i) and (iii) of Proposition 2.5 also holds for nonsymmetric matrices.

PROPOSITION 2.7. *Let A be a tridiagonal nonnegative matrix. Then A is totally positive if and only if its comparison matrix $\mathcal{M}(A)$ is an M-matrix.*

Proof. Let us consider the diagonal matrix $K_n = \operatorname{diag}\{1, -1, 1, \ldots, (-1)^{n+1}\}$. The matrix K_nAK_n has nonnegative diagonal elements and nonpositive off-diagonal elements because A is tridiagonal and nonnegative. Therefore, K_nAK_n is the comparison matrix $\mathcal{M}(A)$ of A. Furthermore, the principal minors of K_nAK_n coincide with the principal minors of A because they only differ in the fact that in K_nAK_n the even rows and even columns of A have been multiplied by (-1).

If A is totally positive, then K_nAK_n has nonnegative principal minors and, by Theorem (4.6) of Chapter 6 of [5], $K_nAK_n = \mathcal{M}(A)$ is an M-matrix.

Finally, if $\mathcal{M}(A)(= K_nAK_n)$ is an M-matrix, then it has nonnegative principal minors by Theorem (4.6) of Chapter 6 of [5]. Consequently, the nonnegative matrix A is totally positive by [1, Theorem 2.3]. □

As a consequence of the previous result we may obtain from Theorem (5.14) of Chapter 7 of [5] a result on the convergence of iterative methods for tridiagonal totally positive matrices. As usual, let $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$ and $-L$ and $-U$ be the strictly lower and strictly upper triangular parts of A, respectively. Thus, $A = D - L - U$. The iteration matrices for the Jacobi and successive over relaxation (SOR) methods are $J = D^{-1}(L + U)$ and $H_\omega = (D - \omega L)^{-1}((1 - \omega)D + \omega U)$, respectively.

COROLLARY 2.8. *Let A be a tridiagonal nonsingular totally positive matrix. Then the Jacobi method is convergent and the SOR method converges whenever*

$$0 < \omega < \frac{2}{1 + \rho(|J|)}.$$

3. On the existence of nonintersecting paths in digraphs. In this section we shall deal with digraphs. We shall show how some properties of the totally positive matrices follow easily from their interpretation in terms of digraphs and, conversely, we shall see that information on the existence of nonintersecting paths with positive weight in a digraph can be obtained from the properties of the totally positive matrices.

Following the notations of [6], let $D = (V, A)$ be a *digraph*. We shall assume that D has no loops or multiple edges. So, the elements of A (i.e., the edges) can be identified with ordered pairs (u, v) with $u, v \in V$, $u \neq v$. A *path* in D is a sequence $\pi = u_1 u_2 \cdots u_n$ of elements of V such that $(u_i, u_{i+1}) \in A$ for $i = 1, \ldots, n-1$ (we say that π goes from u_1 to u_n). We say that D is *locally finite* if, for every $u, v \in V$, there are only a finite number of paths from u to v. Let us observe that a locally finite digraph D must be acyclic. We say that D is *weighted* if there is a function $w : A \to \mathbf{R}$. If $w(u, v) \geq 0$ for all $(u, v) \in A$, then we call D a *nonnegative* digraph.

Let $D = (V, A, w)$ be a locally finite, weighted digraph. For a path $\pi = u_0 u_1 \cdots u_k$ in D we define $w(\pi) := \prod_{i=1}^{k} w(u_{i-1}, u_i)$, and, for $u, v \in V$, $P_D(u, v) := \sum_{\pi} w(\pi)$, where the sum is over all paths π in D going from u to v. By convention, $P_D(u, u) := 1$. Given $\mathbf{u} := (u_1, \ldots, u_r)$, $\mathbf{v} := (v_1, \ldots, v_r) \in V^r$, we let

$$N(\mathbf{u}, \mathbf{v}) := \sum_{(\pi_1, \ldots, \pi_r)} w(\pi_1, \ldots, \pi_r),$$

where $w(\pi_1, \ldots, \pi_r) := \prod_{i=1}^{r} w(\pi_i)$ and where the sum is over all r-tuples $(\pi_1, \ldots, \pi_r)$ of paths from $\mathbf{u}$ to $\mathbf{v}$ (i.e., π_i is a path from u_i to v_i, for $i = 1, \ldots, r$) that are *nonintersecting* (i.e., π_i and π_j have no vertices in common if $i \neq j$). Most classes of plane partitions that are of interest (either by association with the representation theory of the classical groups, or for purely combinatorial reasons) can be interpreted as configurations of nonintersecting paths in a digraph. We say that $\mathbf{u}$ and $\mathbf{v}$ are *compatible* if, for every $\sigma \in S_r \backslash \{\mathrm{Id}\}$ (where S_r is the group of permutations of a set of r elements), there are no r-tuples of paths from $(u_1, \ldots, u_r)$ to $(v_{\sigma(1)}, \ldots, v_{\sigma(r)})$ that are nonintersecting. The proof of the following result can be found in [16, Theorem 1.2] or in [15, Lemma 1].

PROPOSITION 3.1. *Let $D = (V, A, w)$ be a locally finite, weighted digraph and $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ be compatible. Then*

$$N(\mathbf{u}, \mathbf{v}) = \det\left[(P_D(u_i, v_j))_{1 \leq i, j \leq n}\right]. \tag{3.1}$$

The previous result gives no information about which sets of the vertices are compatible. However, if D is planar it is often possible to take advantage of the underlying topology, as shown in [16]. For example, suppose that one may pass a Jordan curve C through two sets of vertices I and J so that all paths from I to J are contained in the interior of C. If the vertices of I and J are arranged along two distinct segments of C, then I must be compatible with J. [16, Proposition 1.4] gives an algebraic method for identifying compatible sets of vertices.

Proposition 3.1 implies that if D is a locally finite, nonnegative digraph and $(u_1, \ldots, u_n)$, $(v_1, \ldots, v_n) \in V^n$ are compatible, then the matrix $(P_D(u_i, v_j))_{1 \leq i, j \leq n}$ has a nonnegative determinant. The following concept leads to totally positive matrices. We say that $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ are *fully compatible* if $(u_{i_1}, \ldots, u_{i_r})$ and $(v_{j_1}, \ldots, v_{j_r})$ are compatible for all $(i_1, \ldots, i_r)$, $(j_1, \ldots, j_r) \in Q_{r,n}$ and $1 \leq r \leq n$. The following characterization of totally positive matrices corresponds to [6, Theorem 3.1].

THEOREM 3.2. *Let U be an $n \times n$ matrix. Then U is totally positive if and only if there exists a planar, locally finite, nonnegative digraph $D = (V, A, w)$ and $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ fully compatible such that $U = (P_D(u_i, v_j))_{1 \leq i,j \leq n}$.*

If $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ are compatible and (3.1) is strictly positive (and so there exists an n-tuple of nonintersecting paths with positive weight from each u_i to each v_i), we say that $\mathbf{u}$ and $\mathbf{v}$ are *strictly compatible*.

The following result follows immediately from the previous definitions.

LEMMA 3.3. *Let $D = (V, A, w)$ be a locally finite, nonnegative digraph and $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ be strictly compatible. If $(u_{i_1}, \ldots, u_{i_k})$ is compatible with $(v_{i_1}, \ldots, v_{i_k})$ ($(i_1, \ldots, i_k) \in Q_{k,n}$, $1 \leq k \leq n$) then they are strictly compatible.*

PROPOSITION 3.4. *Let $D = (V, A, w)$ be a locally finite, nonnegative digraph. Let $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ be strictly compatible and such that for all $(i_1, \ldots, i_k) \in Q_{k,n}$, $1 \leq k \leq n$, $(u_1, \ldots, u_k)$ is compatible with $(v_{i_1}, \ldots, v_{i_k})$ and $(v_1 \ldots, v_k)$ is compatible with $(u_{i_1}, \ldots, u_{i_k})$. Then there exist a locally finite, nonnegative digraph $\bar{D} = (\bar{V}, \bar{A}, \bar{w})$ and $\bar{\mathbf{u}} = (\bar{u}_1, \ldots, \bar{u}_n)$, $\bar{\mathbf{v}} = (\bar{v}_1, \ldots, \bar{v}_n) \in \bar{V}^n$ such that they are fully compatible and $P_D(u_i, v_j) = P_{\bar{D}}(\bar{u}_i, \bar{v}_j)$ for all $i, j \in \{1, \ldots, n\}$.*

Proof. Let $B := (P_D(u_i, v_j))_{1 \leq i,j \leq n}$. By hypothesis, $\det B[1, \ldots, k | i_1, \ldots, i_k] \geq 0$ and $\det B[i_1, \ldots, i_k | 1, \ldots, k] \geq 0$ for all $(i_1, \ldots, i_k) \in Q_{k,n}$, $1 \leq k \leq n$. Besides, for all $k \in \{1 \ldots, n\}$ $(u_1, \ldots, u_k)$ is compatible with $(v_1 \ldots, v_k)$ and, since $\mathbf{u}$ and $\mathbf{v}$ are strictly compatible, $(u_1, \ldots, u_k)$ is strictly compatible with $(v_1 \ldots, v_k)$ by Lemma 3.3. Thus, $\det B[1, \ldots, k] > 0$ for all $k \in \{1 \ldots, n\}$. So, by [10, Theorem 3.1], B is totally positive. Now the result follows from Theorem 3.2. ∎

The next result illustrates how Theorem 3.2 can be used to obtain properties of totally positive matrices. In fact, we shall show that the well-known strict positivity of the principal minors of a nonsingular totally positive matrix (see for instance [1, Corollary 3.8]) is a straightforward consequence of Theorem 3.2.

PROPOSITION 3.5. *If an $n \times n$ totally positive matrix B is nonsingular, then $\det B[\alpha] > 0$ for every $k \in \{1, \ldots, n\}$ and $\alpha \in Q_{k,n}$.*

Proof. By Theorem 3.2 there exists a planar, locally finite, nonnegative digraph $D = (V, A, w)$ and $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ fully compatible satisfying that $B = (P_D(u_i, v_j))_{1 \leq i,j \leq n}$. Since B is nonsingular, we have that $\mathbf{u}$ and $\mathbf{v}$ are in fact strictly compatible. By Lemma 3.3 we have that $(u_{i_1}, \ldots, u_{i_r})$ and $(v_{i_1}, \ldots, v_{i_r})$ are strictly compatible for all $(i_1, \ldots, i_r) \in Q_{r,n}$ and $1 \leq r \leq n$, and then the result follows. ∎

Now we shall apply some results of the theory of totally positive matrices to know the existence of nonintersecting paths with positive weight in a digraph. We say that $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ are *strictly fully compatible* if $(u_{i_1}, \ldots, u_{i_r})$ and $(v_{j_1}, \ldots, v_{j_r})$ are strictly compatible for all $(i_1, \ldots, i_r)$, $(j_1, \ldots, j_r) \in Q_{r,n}$ and $1 \leq r \leq n$ (and so there will be nonintersecting paths with positive weight from $(u_{i_1}, \ldots, u_{i_r})$ to $(v_{j_1}, \ldots, v_{j_r})$). In the next result we shall deduce the existence of r-tuples of paths with positive weight from all sets of vertices $(u_{i_1}, \ldots, u_{i_r})$ to all sets $(v_{j_1}, \ldots, v_{j_r})$ ($(i_1, \ldots, i_r)$, $(j_1, \ldots, j_r) \in Q_{r,n}$ and $1 \leq r \leq n$) from the existence of nonintersecting paths with positive weight between some special sets of vertices.

PROPOSITION 3.6. *Let $D = (V, A, w)$ be a locally finite, nonnegative digraph and $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ be fully compatible. If for $k = 1, 2, \ldots, n$ we have that $(u_1, u_2, \ldots, u_k)$ is strictly compatible with $(v_{n-k+1}, v_{n-k+2}, \ldots, v_n)$ and also that $(u_{n-k+1}, u_{n-k+2}, \ldots, u_n)$ is strictly compatible with $(v_1, v_2, \ldots, v_k)$, then $\mathbf{u}$ and $\mathbf{v}$ are strictly fully compatible.*

Proof. By Proposition 3.1 the matrix $B := (P_D(u_i, v_j))_{1\le i,j\le n}$ is totally positive. By hypothesis, for $k = 1, 2, \ldots, n$ $\det B\,[1, 2, \ldots, k|n-k+1, n-k+2, \ldots, n] > 0$ and $\det B\,[n-k+1, n-k+2, \ldots, n|1, 2, \ldots, k] > 0$. Now the result follows from [9, Theorem 4.3]. □

In the previous result we have proved that **u** and **v** are strictly fully compatible assuming previously that **u** and **v** are fully compatible. In the next result we remove this assumption and we obtain a result of a similar nature to Proposition 3.4.

PROPOSITION 3.7. *Let $D = (V, A, w)$ be a locally finite, nonnegative digraph and $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ be fully compatible. Let us assume that, for $k = 1, 2, \ldots, n$, $(u_1, u_2, \ldots, u_k)$ is strictly compatible with $(v_{n-k+1}, v_{n-k+2}, \ldots, v_n)$ and also that $(u_{n-k+1}, u_{n-k+2}, \ldots, u_n)$ is strictly compatible with $(v_1, v_2, \ldots, v_k)$, and that for all $d \in \{1, 2, \ldots, n-k+1\}$ $(u_1, u_2, \ldots, u_k)$ is compatible with $(v_d, v_{d+1}, \ldots, v_{d+k-1})$ and $(u_d, u_{d+1}, \ldots, u_{d+k-1})$ is compatible with $(v_1, v_2, \ldots, v_k)$. Then there exist a locally finite, nonnegative digraph $\bar{D} = (\bar{V}, \bar{A}, \bar{w})$ and $\bar{\mathbf{u}} = (\bar{u}_1, \ldots, \bar{u}_n)$, $\bar{\mathbf{v}} = (\bar{v}_1, \ldots, \bar{v}_n) \in \bar{V}^n$ such that they are strictly fully compatible and $P_D(u_i, v_j) = P_{\bar{D}}(\bar{u}_i, \bar{v}_j)$ for all $i, j \in \{1, \ldots, n\}$.*

Proof. Let $B := (P_D(u_i, v_j))_{1\le i,j\le n}$. For all $k = 1, 2, \ldots, n$ and for all $d \in \{1, 2, \ldots, n-k+1\}$ $(u_1, u_2, \ldots, u_k)$ is compatible with $(v_d, v_{d+1}, \ldots, v_{d+k-1})$. Since we have that $(u_1, \ldots, u_k, u_{k+1}, \ldots, u_{n-d+1})$ is strictly compatible with $(v_d, \ldots, v_{d+k-1}, v_{d+k}, \ldots, v_n)$, we obtain by Lemma 3.3 that $(u_1, \ldots, u_k)$ is strictly compatible with $(v_d, \ldots, v_{d+k-1})$. Thus, $\det B\,[1, 2, \ldots, k|d, d+1, \ldots, d+k-1] > 0$ for all $k = 1, 2, \ldots, n$ and for all $d \in \{1, 2, \ldots, n-k+1\}$. Analogously the positivity of the determinants $\det B\,[d, d+1, \ldots, d+k-1|1, 2, \ldots, k]$ for all $k = 1, 2, \ldots, n$ and for all $d \in \{1, 2, \ldots, n-k+1\}$ can be proven. Now the result follows from [9, Theorem 4.1] and from Theorem 3.2. □

Given $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$, a (trivial) necessary condition for the existence of nonintersecting paths with positive weight from $(u_{i_1}, \ldots, u_{i_r})$ to $(v_{j_1}, \ldots, v_{j_r})$ is that there exist paths with positive weight from u_{i_l} to v_{j_l} for $l = 1, \ldots, r$. Two compatible **u**,**v** are called *almost strictly fully compatible* if this necessary condition is also sufficient, i.e., $(u_{i_1}, \ldots, u_{i_r})$ is strictly compatible with $(v_{j_1}, \ldots, v_{j_r})$ if and only if there exist paths with positive weight from u_{i_l} to v_{j_l} for $l = 1, \ldots, r$. Let us remark that if **u**,**v** are almost strictly fully compatible (respectively, strictly fully compatible), then the corresponding matrices $B = (P_D(u_i, v_j))_{1\le i,j\le n}$ are almost strictly totally positive (respectively, strictly totally positive). These matrix definitions can be found in [9] and [12].

The next result will provide a sufficient condition to prove that **u**,**v** are almost strictly fully compatible. Previously, we have to recall some well-known facts on the zero pattern of a nonsingular totally positive matrix $A = (a_{ij})_{1\le i,j\le n}$. Taking into account that by Proposition 3.5 it cannot have a zero as a diagonal entry and that all its 2×2 minors are nonnegative, one can deduce that its entries satisfy

$$\begin{aligned} &a_{ij} = 0, i > j \Rightarrow a_{hk} = 0 \quad \forall h \ge i, k \le j, \\ &a_{ij} = 0, i < j \Rightarrow a_{hk} = 0 \quad \forall h \le i, k \ge j. \end{aligned} \tag{3.2}$$

Thus, the patterns of zeros of these matrices are determined by the following indices. For an $n \times n$ matrix A let us denote

$$\begin{aligned} &i_0 = 1, \quad j_0 = 1; \\ &\text{for} \quad t = 1, \ldots, l : \end{aligned}$$

$$i_t = \max\{i|a_{i,j_{t-1}} \neq 0\} + 1 \quad (\leq n+1),$$
$$j_t = \max\{j|a_{i_t,j} = 0\} + 1 \quad (\leq n+1),$$

where l is given in this recurrent definition by $i_l = n+1$. Analogously we denote

$$\begin{aligned} &\hat{j}_0 = 1, \quad \hat{i}_0 = 1; \\ &\quad \text{for} \quad t = 1, \ldots, r: \\ &\qquad \hat{j}_t = \max\{j|a_{\hat{i}_{t-1},j} \neq 0\} + 1, \\ &\qquad \hat{i}_t = \max\{i|a_{i,\hat{j}_t} = 0\} + 1, \end{aligned}$$

where $\hat{j}_r = n+1$. In other words, the entries below the places $(i_1 - 1, j)$ with $j_0 \leq j < j_1$, $(i_2 - 1, j)$ with $j_1 \leq j < j_2$, ..., $(i_{l-1} - 1, j)$ with $j_{l-2} \leq j < j_{l-1}$ are zero. So are the entries to the right of the places $(i, \hat{j}_1 - 1)$ with $\hat{i}_0 \leq i < \hat{i}_1$, $(i, \hat{j}_2 - 1)$ with $\hat{i}_1 \leq i < \hat{i}_2$, ..., $(i, \hat{j}_{r-1} - 1)$ with $\hat{i}_{r-2} \leq i < \hat{i}_{r-1}$. On the other hand, the entries of both lists, those above the first list and those to the left of the last list, are nonzero. We shall say that the matrix A has a zero pattern given by $I = \{i_0, i_1, \ldots, i_l\}$, $J = \{j_0, j_1, \ldots, j_l\}$, $\hat{I} = \{\hat{i}_0, \hat{i}_1, \ldots, \hat{i}_r\}$, and $\hat{J} = \{\hat{j}_0, \hat{j}_1, \ldots, \hat{j}_r\}$. Only matrices with these patterns of zeros and all the other entries positive can be nonsingular totally positive. Besides, we have that

$$\begin{aligned} &i_t > j_t, \quad t = 1, \ldots, l-1, \\ &\hat{j}_t > \hat{i}_t, \quad t = 1, \ldots, r-1. \end{aligned} \tag{3.3}$$

Let us consider an example of a 10×10 matrix with $l = r = 3$ and $\{i_0, i_1, i_2, i_3\} = \{1, 6, 9, 11\}$, $\{j_0, j_1, j_2, j_3\} = \{1, 3, 5, 11\}$, $\{\hat{j}_0, \hat{j}_1, \hat{j}_2, \hat{j}_3\} = \{1, 7, 9, 11\}$, $\{\hat{i}_0, \hat{i}_1, \hat{i}_2, \hat{i}_3\} = \{1, 3, 5, 11\}$. Entries represented by the symbol * are nonzero:

$$\begin{pmatrix} * & * & * & * & * & * & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & 0 & 0 & 0 & 0 \\ * & * & * & * & * & * & * & * & 0 & 0 \\ * & * & * & * & * & * & * & * & 0 & 0 \\ * & * & * & * & * & * & * & * & * & * \\ 0 & 0 & * & * & * & * & * & * & * & * \\ 0 & 0 & * & * & * & * & * & * & * & * \\ 0 & 0 & * & * & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & * & * & * & * & * & * \end{pmatrix}.$$

PROPOSITION 3.8. *Let $D = (V, A, w)$ be a locally finite, nonnegative digraph and $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n) \in V^n$ be fully compatible and strictly compatible. Let $B := (P_D(u_i, v_j))_{1 \leq i,j \leq n}$, with a zero pattern given by $I, J, \hat{I}, \hat{J}$ as above. Let us assume also that for $1 \leq t \leq l$ and $j_{t-1} \leq h < j_t$, $(u_{i_t - 1 - h + j_k}, \ldots, u_{i_t - 1})$ is strictly compatible with $(v_{j_k}, v_{j_k+1}, \ldots, v_h)$ (where $j_k = \max\{j_s | s \leq t-1, h - j_s < i_t - i_s\}$) and that for $1 \leq t \leq r$ and $\hat{i}_{t-1} \leq h < \hat{i}_t$, $(u_{\hat{i}_k}, u_{\hat{i}_k+1} \ldots, u_h)$ is strictly compatible with $(v_{\hat{j}_t - 1 - h + \hat{i}_k}, \ldots, v_{\hat{j}_t - 1})$ (where $\hat{i}_k = \max\{\hat{i}_s | s \leq t-1, h - \hat{i}_s < \hat{j}_t - \hat{j}_s\}$). Then $\mathbf{u}$ and $\mathbf{v}$ are almost strictly fully compatible.*

Proof. Since $\mathbf{u}$ and $\mathbf{v}$ are fully compatible, the matrix B is totally positive by Theorem 3.2. B is also nonsingular because $\mathbf{u}$ and $\mathbf{v}$ are strictly compatible. Thus, B has a zero pattern given by $I, J, \hat{I}, \hat{J}$ as above. Now the result follows easily from the equivalence of (1) and (3) in [12, Theorem 3.3]. ∎

Finally, let us mention that [9, Theorem 4.3 (ii)] (respectively, [12, Theorem 3.3 (2)]) provides an algorithmic way to check that fully compatible sets of vertices $\mathbf{u} = (u_1, \ldots, u_n)$, $\mathbf{v} = (v_1, \ldots, v_n)$ are strictly fully compatible (respectively, almost strictly fully compatible): we have to check the positivity of the pivots (respectively, of the pivots corresponding to the nonzero elements) when we perform the Neville elimination of the matrix $(P_D(u_i, v_j))_{1 \leq i,j \leq n}$. Roughly speaking, *Neville elimination* is a procedure to create zeros in a matrix by means of adding to a given row a suitable multiple of the previous one. In [9] there appears a detailed exposition of this elimination procedure.

REFERENCES

[1] T. Ando, *Totally positive matrices*, Linear Algebra Appl., 90 (1987), pp. 165–219.

[2] A. Berman, *Complete positivity*, Linear Algebra Appl., 107 (1988), pp. 57–63.

[3] A. Berman, *Completely positive graphs*, in Combinatorial and Graph-Theoretical Problems in Linear Algebra, R. A. Brualdi, S. Friedland, and V. Klee, eds., IMA Vol. Math. Appl. 50, Springer-Verlag, New York, 1993, pp. 229–233.

[4] A. Berman and N. Kogan, *Characterization of completely positive graphs*, Discrete Math., 114 (1993), pp. 297–304.

[5] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics App. Math. 9, SIAM, Philadelphia, PA, 1996.

[6] F. Brenti, *Combinatorics and total positivity*, J. Combin. Theory Ser. A, 71 (1995), pp. 175–218.

[7] R. A. Brualdi and H. J. Ryser, *Combinatorial Matrix Theory*, Encyclopedia Math. Appl. 39, Cambridge University Press, Cambridge, 1991.

[8] J. H. Drew, C. R. Johnson, and R. Loewy, *Completely positive matrices associated with M-matrices*, Linear and Multilinear Algebra, 37 (1994), pp. 303–310.

[9] M. Gasca and J. M. Peña, *Total positivity and Neville elimination*, Linear Algebra Appl., 165 (1992), pp. 25–44.

[10] M. Gasca and J. M. Peña, *Total positivity, QR factorization and Neville elimination*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1132–1140.

[11] M. Gasca and J. M. Peña, *A matricial description of Neville elimination with applications to total positivity*, Linear Algebra Appl., 202 (1994), pp. 33–54.

[12] M. Gasca and J. M. Peña, *On the characterization of almost strictly totally positive matrices*, Adv. Comput. Math., 3 (1995), pp. 239–250.

[13] C. R. Johnson, D. D. Olesky, and P. van den Driessche, *A class of M-matrices with tree graphs*, SIAM J. Alg. Disc. Math., 4 (1983), pp. 476–480.

[14] S. Karlin, *Total Positivity*, Stanford University Press, Stanford, CA, 1968.

[15] B. Lindström, *On the vector representations of induced matroids*, Bull. London Math. Soc., 5 (1973), pp. 85–90.

[16] J. R. Stembridge, *Non-intersecting paths, pfaffians, and plane partitions*, Adv. Math., 83 (1990), pp. 96–131.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 378–406, April 1998

007

THE QR DECOMPOSITION AND THE SINGULAR VALUE DECOMPOSITION IN THE SYMMETRIZED MAX-PLUS ALGEBRA*

BART DE SCHUTTER† AND BART DE MOOR†

Abstract. In this paper we discuss matrix decompositions in the symmetrized max-plus algebra. The max-plus algebra has maximization and addition as basic operations. In contrast to linear algebra, many fundamental problems in the max-plus algebra still have to be solved. In this paper we discuss max-algebraic analogues of some basic matrix decompositions from linear algebra. We show that we can use algorithms from linear algebra to prove the existence of max-algebraic analogues of the QR decomposition, the singular value decomposition (SVD), the Hessenberg decomposition, the LU decomposition, and so on.

Key words. max-plus algebra, matrix decompositions, QR decomposition, singular value decomposition

AMS subject classifications. 15A23, 16Y99

PII. S0895479896304782

1. Introduction.

In recent years both industry and the academic world have become more and more interested in techniques to model, analyze, and control complex systems such as flexible manufacturing systems, telecommunication networks, parallel processing systems, traffic control systems, logistic systems, and so on. These systems are typical examples of *discrete event systems* (DESs), the subject of an emerging discipline in system and control theory. The class of the DESs essentially contains man-made systems that consist of a finite number of resources (e.g., machines, communications channels, or processors) that are shared by several users (e.g., product types, information packets, or jobs), all of which contribute to the achievement of some common goal (e.g., the assembly of products, the end-to-end transmission of a set of information packets, or a parallel computation). Although in general DESs lead to a nonlinear description in conventional algebra, there exists a subclass of DESs for which this model becomes "linear" when we formulate it in the max-plus algebra [1, 5]. DESs that belong to this subclass are called *max-linear DESs*.

The basic operations of the max-plus algebra are maximization and addition. There exists a remarkable analogy between the basic operations of the max-plus algebra on the one hand, and the basic operations of conventional algebra (addition and multiplication) on the other hand. As a consequence many concepts and properties of conventional algebra (such as the Cayley–Hamilton theorem, eigenvectors, eigenvalues, and Cramer's rule) also have a max-algebraic analogue. This analogy also allows us to translate many concepts, properties, and techniques from conventional linear system theory to system theory for max-linear DESs. However, there are also some major differences that prevent a straightforward translation of properties, concepts, and algorithms from conventional linear algebra and linear system theory to max-plus algebra and max-algebraic system theory for DESs.

Compared to linear algebra and linear system theory, the max-plus algebra and the max-algebraic system theory for DESs is at present far from fully developed, and

*Received by the editors June 5, 1996; accepted for publication (in revised form) by V. Mehrmann March 11, 1997.

http://www.siam.org/journals/simax/19-2/30478.html

†ESAT/SISTA, Kardinaal Mercierlaan 94, B-3001 Heverlee (Leuven), Belgium (bart.deschutter@esat.kuleuven.ac.be, bart.demoor@esat.kuleuven.ac.be).

much research on this topic is still needed in order to get a complete system theory. The main goal of this paper is to fill one of the gaps in the theory of the max-plus algebra by showing that there exist max-algebraic analogues of many fundamental matrix decompositions from linear algebra such as the QR decomposition and the singular value decomposition (SVD). These matrix decompositions are important tools in many linear algebra algorithms (see [14] and the references cited therein) and in many contemporary algorithms for the identification of linear systems (see [21, 22, 33, 34, 35] and the references cited therein).

In [30], Olsder and Roos have used asymptotic equivalences to show that every matrix has at least one max-algebraic eigenvalue and to prove a max-algebraic version of Cramer's rule and of the Cayley–Hamilton theorem. We shall use an extended and formalized version of their technique to prove the existence of the QR decomposition and the SVD in the symmetrized max-plus algebra. In our existence proof we shall use algorithms from linear algebra. This proof technique can easily be adapted to prove the existence of max-algebraic analogues of many other matrix decompositions from linear algebra such as the Hessenberg decomposition, the LU decomposition, the eigenvalue decomposition, the Schur decomposition, and so on.

This paper is organized as follows. After introducing some concepts and definitions in section 2, we give a short introduction to the max-plus algebra and the symmetrized max-plus algebra in section 3. Next we establish a link between a ring of real functions (with conventional addition and multiplication as basic operations) and the symmetrized max-plus algebra. In section 5 we use this link to define the QR decomposition and the SVD of a matrix in the symmetrized max-plus algebra and to prove the existence of these decompositions. We conclude with an example.

2. Notations and definitions. In this section we give some definitions that will be used in the next sections.

The set of all reals except for zero is represented by $\mathbb{R}_0$ ($\mathbb{R}_0 = \mathbb{R} \setminus \{0\}$). The set of nonnegative real numbers is denoted by $\mathbb{R}^+$, and the set of nonpositive real numbers is denoted by $\mathbb{R}^-$. We have $\mathbb{R}_0^+ = \mathbb{R}^+ \setminus \{0\}$.

We shall use "vector" as a synonym for "n-tuple." Furthermore, all vectors are assumed to be column vectors. If a is a vector, then a_i is the ith component of a. If A is a matrix, then a_{ij} or $(A)_{ij}$ is the entry on the ith row and the jth column. The $n \times n$ identity matrix is denoted by I_n and the $m \times n$ zero matrix is denoted by $O_{m\times n}$.

The matrix $A \in \mathbb{R}^{n\times n}$ is called orthogonal if $A^T A = I_n$.

The Frobenius norm of the matrix $A \in \mathbb{R}^{m\times n}$ is represented by

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} \quad .$$

The 2-norm of the vector a is defined by $\|a\|_2 = \sqrt{a^T a}$ and the 2-norm of the matrix A is defined by

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 \quad .$$

If $A \in \mathbb{R}^{m\times n}$, then there exist an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ and an upper triangular matrix $R \in \mathbb{R}^{m\times n}$ such that $A = QR$. We say that QR is a *QR decomposition* of A.

Let $A \in \mathbb{R}^{m\times n}$ and let $r = \min(m, n)$. Then there exist a diagonal matrix $\Sigma \in \mathbb{R}^{m\times n}$ and two orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that

$$A = U\,\Sigma\,V^T \tag{1}$$

with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$, where $\sigma_i = (\Sigma)_{ii}$ for $i = 1, 2, \ldots, r$. Factorization (1) is called an SVD of A. The diagonal entries of Σ are the singular values of A. We have $\sigma_1 = \|A\|_2$. The columns of U are the left singular vectors of A and the columns of V are the right singular vectors of A. For more information on the QR decomposition and the SVD the interested reader is referred to [14, 18].

We use f, $f(\cdot)$ or $x \mapsto f(x)$ to represent a function. The domain of definition of the function f is denoted by $\operatorname{dom} f$ and the value of f at $x \in \operatorname{dom} f$ is denoted by $f(x)$.

DEFINITION 2.1 (analytic function). *Let f be a real function and let $\alpha \in \mathbb{R}$ be an interior point of* $\operatorname{dom} f$. *Then f is analytic in α if the Taylor series of f with center α exists and if there is a neighborhood of α where this Taylor series converges to f.*

A real function f is analytic in an interval $[\alpha, \beta] \subseteq \operatorname{dom} f$ if it is analytic in every point of that interval.

A real matrix-valued function $\tilde{F}$ is analytic in $[\alpha, \beta] \subseteq \operatorname{dom} \tilde{F}$ if all its entries are analytic in $[\alpha, \beta]$.

DEFINITION 2.2 (asymptotic equivalence in the neighborhood of ∞). *Let f and g be real functions such that ∞ is an accumulation point of* $\operatorname{dom} f$ *and* $\operatorname{dom} g$.

If there is no real number K such that g is identically zero in $[K, \infty)$, then we say that f is asymptotically equivalent to g in the neighborhood of ∞, denoted by $f(x) \sim g(x)$, $x \to \infty$, if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$.

If there exists a real number L such that both f and g are identically zero in $[L, \infty)$, then we also say that $f(x) \sim g(x)$, $x \to \infty$.

Let $\tilde{F}$ and $\tilde{G}$ be real $m \times n$ matrix-valued functions such that ∞ is an accumulation point of $\operatorname{dom} \tilde{F}$ *and* $\operatorname{dom} \tilde{G}$. *Then $\tilde{F}(x) \sim \tilde{G}(x)$, $x \to \infty$ if $\tilde{f}_{ij}(x) \sim \tilde{g}_{ij}(x)$, $x \to \infty$ for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.*

The main difference with the conventional definition of asymptotic equivalence is that Definition 2.2 also allows us to say that a function is asymptotically equivalent to zero in the neighborhood of ∞: $f(x) \sim 0$, $x \to \infty$ if there exists a real number L such that $f(x) = 0$ for all $x \geq L$.

3. The max-plus algebra and the symmetrized max-plus algebra. In this section we give a short introduction to the max-plus algebra and the symmetrized max-plus algebra. A complete overview of the max-plus algebra can be found in [1, 5, 12].

3.1. The max-plus algebra. The basic max-algebraic operations are defined as follows:

$$x \oplus y = \max(x, y), \tag{2}$$

$$x \otimes y = x + y \tag{3}$$

for $x, y \in \mathbb{R} \cup \{-\infty\}$ with, by definition, $\max(x, -\infty) = x$ and $x + (-\infty) = -\infty$ for all $x \in \mathbb{R} \cup \{-\infty\}$. The reason for using the symbols $\oplus$ and $\otimes$ to represent maximization and addition is that there is a remarkable analogy between $\oplus$ and addition, and between $\otimes$ and multiplication: many concepts and properties from conventional linear algebra (such as the Cayley–Hamilton theorem, eigenvectors, eigenvalues, and Cramer's rule) can be translated to the (symmetrized) max-plus algebra by replacing

$+$ by $\oplus$ and $\times$ by $\otimes$ (see also section 4). Therefore, we also call $\oplus$ the max-algebraic addition. Likewise, we call $\otimes$ the max-algebraic multiplication. The resulting algebraic structure $\mathbb{R}_{\max} = (\mathbb{R} \cup \{-\infty\}, \oplus, \otimes)$ is called the *max-plus algebra*. Define $\mathbb{R}_\varepsilon = \mathbb{R} \cup \{-\infty\}$. The zero element for $\oplus$ in $\mathbb{R}_\varepsilon$ is represented by $\varepsilon \stackrel{\text{def}}{=} -\infty$. So $x \oplus \varepsilon = x = \varepsilon \oplus x$ for all $x \in \mathbb{R}_\varepsilon$.

Let $r \in \mathbb{R}$. The rth max-algebraic power of $x \in \mathbb{R}$ is denoted by $x^{\otimes^r}$ and corresponds to rx in conventional algebra. If $x \in \mathbb{R}$, then $x^{\otimes^0} = 0$ and the inverse element of x with respect to $\otimes$ is $x^{\otimes^{-1}} = -x$. There is no inverse element for ε since ε is absorbing for $\otimes$. If $r > 0$, then $\varepsilon^{\otimes^r} = \varepsilon$. If $r \leq 0$, then $\varepsilon^{\otimes^r}$ is not defined.

The rules for the order of evaluation of the max-algebraic operators are similar to those of conventional algebra. So max-algebraic power has the highest priority, and max-algebraic multiplication has a higher priority than max-algebraic addition.

Consider the finite sequence $a_1, a_2, \ldots, a_n$ with $a_i \in \mathbb{R}_\varepsilon$ for all i. We define

$$\bigoplus_{i=1}^{n} a_i \;=\; a_1 \;\oplus\; a_2 \;\oplus\; \cdots \;\oplus\; a_n.$$

The matrix E_n is the $n \times n$ max-algebraic identity matrix:

$$\begin{aligned} &(E_n)_{ii} = 0 \qquad \text{for } i = 1, 2, \ldots, n\,, \\ &(E_n)_{ij} = \varepsilon \qquad \text{for } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, n \text{ with } i \neq j\,. \end{aligned}$$

The $m \times n$ max-algebraic zero matrix is represented by $\mathcal{E}_{m\times n}$: we have $(\mathcal{E}_{m\times n})_{ij} = \varepsilon$ for all i, j.

The off-diagonal entries of a max-algebraic diagonal matrix $D \in \mathbb{R}_\varepsilon^{m\times n}$ are equal to ε: $d_{ij} = \varepsilon$ for all i, j with $i \neq j$. A matrix $R \in \mathbb{R}_\varepsilon^{m\times n}$ is a max-algebraic upper triangular matrix if $r_{ij} = \varepsilon$ for all i, j with $i > j$. If we permute the rows or the columns of the max-algebraic identity matrix, we obtain a max-algebraic permutation matrix.

The operations $\oplus$ and $\otimes$ are extended to matrices as follows. If $\alpha \in \mathbb{R}_\varepsilon$ and if $A, B \in \mathbb{R}_\varepsilon^{m\times n}$, then

$$(\alpha \otimes A)_{ij} \;=\; \alpha \otimes a_{ij} \qquad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n$$

and

$$(A \oplus B)_{ij} \;=\; a_{ij} \oplus b_{ij} \qquad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n\,.$$

If $A \in \mathbb{R}_\varepsilon^{m\times p}$ and $B \in \mathbb{R}_\varepsilon^{p\times n}$, then

$$(A \otimes B)_{ij} \;=\; \bigoplus_{k=1}^{p} a_{ik} \otimes b_{kj} \qquad \text{for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n\,.$$

3.2. The symmetrized max-plus algebra. One of the major differences between conventional algebra and the max-plus algebra is that there exist no inverse elements with respect to $\oplus$ in $\mathbb{R}_\varepsilon$: if $x \in \mathbb{R}_\varepsilon$, then there does not exist an element $y_x \in \mathbb{R}_\varepsilon$ such that $x \oplus y_x = \varepsilon = y_x \oplus x$, except when x is equal to ε. So $(\mathbb{R}_\varepsilon, \oplus)$ is not a group. Therefore, we now introduce $\mathbb{S}_{\max}$ [1, 12, 25], which is a kind of symmetrization of the max-plus algebra. This can be compared with the extension of $(\mathbb{N}, +, \times)$ to $(\mathbb{Z}, +, \times)$. In section 4 we shall show that $\mathbb{R}_{\max}$ corresponds to a set of

nonnegative real functions with addition and multiplication as basic operations and that $\mathbb{S}_{\max}$ corresponds to a set of real functions with addition and multiplication as basic operations. Since the $\oplus$ operation is idempotent, we cannot use the conventional symmetrization technique since every idempotent group reduces to a trivial group [1, 25]. Nevertheless, it is possible to adapt the method of the construction of $\mathbb{Z}$ from $\mathbb{N}$ to obtain "balancing" elements rather than inverse elements.

We shall restrict ourselves to a short introduction to the most important features of $\mathbb{S}_{\max}$. This introduction is based on [1, 12, 25]. First we introduce the "algebra of pairs." We consider the set of ordered pairs $\mathcal{P}_\varepsilon \stackrel{\text{def}}{=} \mathbb{R}_\varepsilon \times \mathbb{R}_\varepsilon$ with operations $\oplus$ and $\otimes$ that are defined as follows:

$$(x, y) \oplus (w, z) = (x \oplus w,\ y \oplus z), \tag{4}$$

$$(x, y) \otimes (w, z) = (x \otimes w \oplus y \otimes z,\ x \otimes z \oplus y \otimes w) \tag{5}$$

for (x, y), $(w, z) \in \mathcal{P}_\varepsilon$, and where the operations $\oplus$ and $\otimes$ on the right-hand side correspond to maximization and addition as defined in (2) and (3). The reason for also using $\oplus$ and $\otimes$ on the left-hand side is that these operations correspond to $\oplus$ and $\otimes$ as defined in $\mathbb{R}_{\max}$ as we shall see later on. It is easy to verify the following: in $\mathcal{P}_\varepsilon$ the $\oplus$ operation is associative, commutative, and idempotent, and its zero element is $(\varepsilon, \varepsilon)$; the $\otimes$ operation is associative, commutative, and distributive with respect to $\oplus$; the identity element of $\otimes$ is $(0, \varepsilon)$; and the zero element $(\varepsilon, \varepsilon)$ is absorbing for $\otimes$. We call the structure $(\mathcal{P}_\varepsilon, \oplus, \otimes)$ the *algebra of pairs*.

If $u = (x, y) \in \mathcal{P}_\varepsilon$, then we define the max-absolute value of u as $|u|_\oplus = x \oplus y$ and we introduce two unary operators $\ominus$ (the max-algebraic minus operator) and $(\cdot)^\bullet$ (the balance operator) such that $\ominus u = (y, x)$ and $u^\bullet = u \oplus (\ominus u) = (|u|_\oplus, |u|_\oplus)$. We have

$$u^\bullet = (\ominus u)^\bullet = (u^\bullet)^\bullet, \tag{6}$$

$$u \otimes v^\bullet = (u \otimes v)^\bullet, \tag{7}$$

$$\ominus(\ominus u) = u, \tag{8}$$

$$\ominus(u \oplus v) = (\ominus u) \oplus (\ominus v), \tag{9}$$

$$\ominus(u \otimes v) = (\ominus u) \otimes v \tag{10}$$

for all $u, v \in \mathcal{P}_\varepsilon$. The last three properties allow us to write $u \ominus v$ instead of $u \oplus (\ominus v)$. Since the properties (8)–(10) resemble properties of the $-$ operator in conventional algebra, we could say that the $\ominus$ operator of the algebra of pairs can be considered as the analogue of the $-$ operator of conventional algebra (see also section 4). As for the order of evaluation of the max-algebraic operators, the max-algebraic minus operator has the same, i.e., the lowest, priority as the max-algebraic addition operator.

In conventional algebra we have $x - x = 0$ for all $x \in \mathbb{R}$, but in the algebra of pairs we have $u \ominus u = u^\bullet \neq (\varepsilon, \varepsilon)$ for all $u \in \mathcal{P}_\varepsilon$ unless u is equal to $(\varepsilon, \varepsilon)$, the zero element for $\oplus$ in $\mathcal{P}_\varepsilon$. Therefore, we introduce a new relation.

DEFINITION 3.1 (balance relation). *Consider $u = (x, y)$, $v = (w, z) \in \mathcal{P}_\varepsilon$. We say that u balances v, denoted by $u \nabla v$, if $x \oplus z = y \oplus w$.*

We have $u \ominus u = u^\bullet = (|u|_\oplus, |u|_\oplus) \nabla (\varepsilon, \varepsilon)$ for all $u \in \mathcal{P}_\varepsilon$. The balance relation is reflexive and symmetric, but it is not transitive since, e.g., $(2, 1) \nabla (2, 2)$ and $(2, 2) \nabla (1, 2)$ but $(2, 1) \not\nabla (1, 2)$. Hence, the balance relation is not an equivalence relation and we cannot use it to define the quotient set of $\mathcal{P}_\varepsilon$ by ∇ (as opposed to conventional algebra where $(\mathbb{N} \times \mathbb{N})/=$ yields $\mathbb{Z}$). Therefore, we introduce another

relation that is closely related to the balance relation and that is defined as follows: if (x, y), $(w, z) \in \mathcal{P}_\varepsilon$, then

$$(x,y)\mathcal{B}(w,z) \quad \text{if} \quad \begin{cases} (x,y) \nabla (w,z) & \text{if } x \neq y \text{ and } w \neq z\,, \\ (x,y) = (w,z) & \text{otherwise.} \end{cases}$$

Note that if $u \in \mathcal{P}_\varepsilon$, then we have $u \ominus u \not\mathcal{B} (\varepsilon, \varepsilon)$ unless u is equal to $(\varepsilon, \varepsilon)$. It is easy to verify that $\mathcal{B}$ is an equivalence relation that is compatible with $\oplus$ and $\otimes$, with the balance relation ∇, and with the $\ominus$, $|\cdot|_\oplus$ and $(\cdot)^\bullet$ operators. We can distinguish between three kinds of equivalence classes generated by $\mathcal{B}$:

1. $\overline{(w, -\infty)} = \{\, (w, x) \in \mathcal{P}_\varepsilon \,|\, x < w \,\}$, called max-positive;
2. $\overline{(-\infty, w)} = \{\, (x, w) \in \mathcal{P}_\varepsilon \,|\, x < w \,\}$, called max-negative;
3. $\overline{(w, w)} = \{\, (w, w) \in \mathcal{P}_\varepsilon \,\}$, called balanced.

The class $\overline{(\varepsilon, \varepsilon)}$ is called the max-zero class.

Now we define the quotient set $\mathbb{S} = \mathcal{P}_\varepsilon/\mathcal{B}$. The algebraic structure $\mathbb{S}_{\max} = (\mathbb{S}, \oplus, \otimes)$ is called the *symmetrized max-plus algebra*. By associating $\overline{(w, -\infty)}$ with $w \in \mathbb{R}_\varepsilon$, we can identify $\mathbb{R}_\varepsilon$ with the set of max-positive or max-zero classes denoted by $\mathbb{S}^\oplus$. The set of max-negative or max-zero classes will be denoted by $\mathbb{S}^\ominus$, and the set of balanced classes will be represented by $\mathbb{S}^\bullet$. This results in the following decomposition: $\mathbb{S} = \mathbb{S}^\oplus \cup \mathbb{S}^\ominus \cup \mathbb{S}^\bullet$. Note that the max-zero class $\overline{(\varepsilon, \varepsilon)}$ corresponds to ε. The max-positive elements, the max-negative elements, and ε are called signed. Define $\mathbb{S}^\vee = \mathbb{S}^\oplus \cup \mathbb{S}^\ominus$. Note that $\mathbb{S}^\oplus \cap \mathbb{S}^\ominus \cap \mathbb{S}^\bullet = \{\, \overline{(\varepsilon, \varepsilon)} \,\}$ and $\varepsilon = \ominus\varepsilon = \varepsilon^\bullet$.

These notations allow us to write, e.g., $2 \oplus (\ominus 4)$ instead of $\overline{(2, -\infty)} \oplus \overline{(-\infty, 4)}$. Since $\overline{(2, -\infty)} \oplus \overline{(-\infty, 4)} = \overline{(2, 4)} = \overline{(-\infty, 4)}$, we have $2 \oplus (\ominus 4) = \ominus 4$.

Let $x, y \in \mathbb{R}_\varepsilon$. Since we have

$$(x, -\infty) \oplus (y, -\infty) = (x \oplus y,\, -\infty),$$
$$(x, -\infty) \otimes (y, -\infty) = (x \otimes y,\, -\infty),$$

the operations $\oplus$ and $\otimes$ of the algebra of pairs as defined by (4)–(5) correspond to the operations $\oplus$ and $\otimes$ of the max-plus algebra as defined by (2)–(3).

In general, if $x, y \in \mathbb{R}_\varepsilon$, then we have

$$x \oplus (\ominus y) = x \qquad \text{if } x > y\,, \tag{11}$$
$$x \oplus (\ominus y) = \ominus y \qquad \text{if } x < y\,, \tag{12}$$
$$x \oplus (\ominus x) = x^\bullet\,. \tag{13}$$

Now we give some extra properties of balances that will be used in the next sections.

An element with a $\ominus$ sign can be transferred to the other side of a balance as follows.

PROPOSITION 3.2. *$\forall a, b, c \in \mathbb{S} : a \ominus c \nabla b$ if and only if $a \nabla b \oplus c$.*

If both sides of a balance are signed, we may replace the balance by an equality.

PROPOSITION 3.3. *$\forall a, b \in \mathbb{S}^\vee : a \nabla b \Rightarrow a = b$.*

Let $a \in \mathbb{S}$. The max-positive part $a^\oplus$ and the max-negative part $a^\ominus$ of a are defined as follows:

- if $a \in \mathbb{S}^\oplus$, then $a^\oplus = a$ and $a^\ominus = \varepsilon$;
- if $a \in \mathbb{S}^\ominus$, then $a^\oplus = \varepsilon$ and $a^\ominus = \ominus a$;
- if $a \in \mathbb{S}^\bullet$, then there exists a number $x \in \mathbb{R}_\varepsilon$ such that $a = x^\bullet$ and then $a^\oplus = a^\ominus = x$.

So $a = a^\oplus \ominus a^\ominus$ and $a^\oplus, a^\ominus \in \mathbb{R}_\varepsilon$. Note that a decomposition of the form $a = x \ominus y$ with $x, y \in \mathbb{R}_\varepsilon$ is unique if it is required that either $x \neq \varepsilon$ and $y = \varepsilon$, $x = \varepsilon$ and $y \neq \varepsilon$, or $x = y$. Hence, the decomposition $a = a^\oplus \ominus a^\ominus$ is unique.

Note that $|a|_\oplus = a^\oplus \oplus a^\ominus$ for all $a \in \mathbb{S}$. We say that $a \in \mathbb{S}$ is *finite* if $|a|_\oplus \in \mathbb{R}$. If $|a|_\oplus = \varepsilon$, then we say that a is *infinite*.

Definition 3.1 can now be reformulated as follows.

PROPOSITION 3.4. *$\forall a, b \in \mathbb{S} : a \nabla b$ if and only if $a^\oplus \oplus b^\ominus = a^\ominus \oplus b^\oplus$.*

The balance relation is extended to matrices in the usual way: if $A, B \in \mathbb{S}^{m\times n}$, then $A \nabla B$ if $a_{ij} \nabla b_{ij}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Propositions 3.2 and 3.3 can be extended to the matrix case as follows.

PROPOSITION 3.5. *$\forall A, B, C \in \mathbb{S}^{m\times n} : A \ominus C \nabla B$ if and only if $A \nabla B \oplus C$.*

PROPOSITION 3.6. *$\forall A, B \in (\mathbb{S}^\vee)^{m\times n} : A \nabla B \Rightarrow A = B$.*

We conclude this section with a few extra examples that illustrate the concepts defined above.

Example 3.7. We have $2^\oplus = 2$, $2^\ominus = \varepsilon$, and $(5^\bullet)^\oplus = (5^\bullet)^\ominus = 5$. Hence, $2 \nabla 5^\bullet$ since $2^\oplus \oplus (5^\bullet)^\ominus = 2 \oplus 5 = 5 = \varepsilon \oplus 5 = 2^\ominus \oplus (5^\bullet)^\oplus$.

We have $2 \not\nabla \ominus 5$ since $2^\oplus \oplus (\ominus 5)^\ominus = 2 \oplus 5 = 5 \neq \varepsilon = \varepsilon \oplus \varepsilon = 2^\ominus \oplus (\ominus 5)^\oplus$. □

Example 3.8. Consider the balance $x \oplus 2 \nabla 5$. From Proposition 3.2 it follows that this balance can be rewritten as $x \nabla 5 \ominus 2$ or $x \nabla 5$ since $5 \ominus 2 = 5$ by (11).

If we want a signed solution, the balance $x \nabla 5$ becomes an equality by Proposition 3.3. This yields $x = 5$.

The balanced solutions of $x \nabla 5$ are of the form $x = t^\bullet$ with $t \in \mathbb{R}_\varepsilon$. We have $t^\bullet \nabla 5$ or equivalently $t = 5 \oplus t$ if and only if $t \geq 5$.

So the solution set of $x \oplus 2 \nabla 5$ is given by $\{5\} \cup \{t^\bullet \mid t \in \mathbb{R}_\varepsilon, t \geq 5\}$. □

DEFINITION 3.9 (max-algebraic norm). *Let $a \in \mathbb{S}^n$. The max-algebraic norm of a is defined by*

$$\|a\|_\oplus = \bigoplus_{i=1}^{n} |a_i|_\oplus .$$

The max-algebraic norm of the matrix $A \in \mathbb{S}^{m\times n}$ is defined by

$$\|A\|_\oplus = \bigoplus_{i=1}^{m} \bigoplus_{j=1}^{n} |a_{ij}|_\oplus .$$

The max-algebraic vector norm corresponds to the p-norms from linear algebra since

$$\|a\|_\oplus = \left(\bigoplus_{i=1}^{n} |a_i|_\oplus^{\otimes p} \right)^{\otimes \frac{1}{p}} \qquad \text{for every } a \in \mathbb{S}^n \text{ and every } p \in \mathbb{N}_0 .$$

The max-algebraic matrix norm corresponds to both the Frobenius norm and the p-norms from linear algebra since we have

$$\|A\|_\oplus = \left(\bigoplus_{i=1}^{m} \bigoplus_{j=1}^{n} |a_{ij}|_\oplus^{\otimes 2} \right)^{\otimes \frac{1}{2}} \qquad \text{for every } A \in \mathbb{S}^{m\times n}$$

and also $\|A\|_\oplus = \max_{\|x\|_\oplus = 0} \|A \otimes x\|_\oplus$ (the maximum is reached for $x = O_{n\times 1}$).

4. A link between conventional algebra and the symmetrized max-plus algebra. In [30] Olsder and Roos have used a kind of link between conventional algebra and the max-plus algebra based on asymptotic equivalences to show that every matrix has at least one max-algebraic eigenvalue and to prove a max-algebraic version of Cramer's rule and of the Cayley–Hamilton theorem. In [10] we have extended and formalized this link. Now we recapitulate the reasoning of [10] but in a slightly different form that is mathematically more rigorous.

In the next section we shall encounter functions that are asymptotically equivalent to an exponential of the form νe^{xs} for $s \to \infty$. Since we want to allow exponents that are equal to ε, we set $e^{\varepsilon s}$ equal to zero for all positive real values of s by definition. We also define the following classes of functions:

$$\mathcal{R}_{\mathrm{e}}^{+} = \left\{ f : \mathbb{R}_0^+ \to \mathbb{R}^+ \;\middle|\; f(s) = \sum_{i=0}^{n} \mu_i e^{x_i s} \text{ with } n \in \mathbb{N}, \right.$$
$$\left. \mu_i \in \mathbb{R}_0^+, \text{ and } x_i \in \mathbb{R}_\varepsilon \text{ for all } i \right\},$$
$$\mathcal{R}_{\mathrm{e}} = \left\{ f : \mathbb{R}_0^+ \to \mathbb{R} \;\middle|\; f(s) = \sum_{i=0}^{n} \nu_i e^{x_i s} \text{ with } n \in \mathbb{N}, \right.$$
$$\left. \nu_i \in \mathbb{R}_0, \text{ and } x_i \in \mathbb{R}_\varepsilon \text{ for all } i \right\}.$$

It is easy to verify that $(\mathcal{R}_{\mathrm{e}}, +, \times)$ is a ring.

For all $x, y, z \in \mathbb{R}_\varepsilon$ we have

$$x \oplus y = z \quad \Leftrightarrow \quad e^{xs} + e^{ys} \;\sim\; c e^{zs}\,,\; s \to \infty, \tag{14}$$

$$x \otimes y = z \quad \Leftrightarrow \quad e^{xs} \cdot e^{ys} = e^{zs} \qquad \text{for all } s \in \mathbb{R}_0^+, \tag{15}$$

where $c = 1$ if $x \neq y$ and $c = 2$ if $x = y$. The relations (14) and (15) show that there exists a connection between the operations $\oplus$ and $\otimes$ performed on the real numbers and $-\infty$, and the operations $+$ and $\times$ performed on exponentials. We shall extend this link between $(\mathcal{R}_{\mathrm{e}}^{+}, +, \times)$ and $\mathbb{R}_{\max}$ that has already been used in [26, 27, 28, 29, 30]—and under a slightly different form in [6]—to $\mathbb{S}_{\max}$.

We define a mapping $\mathcal{F}$ with domain of definition $\mathbb{S} \times \mathbb{R}_0 \times \mathbb{R}_0^+$ and with

$$\begin{aligned} \mathcal{F}(a, \mu, s) &= |\mu| e^{as} && \text{if } a \in \mathbb{S}^{\oplus}, \\ \mathcal{F}(a, \mu, s) &= -|\mu| e^{|a|_{\oplus} s} && \text{if } a \in \mathbb{S}^{\ominus}, \\ \mathcal{F}(a, \mu, s) &= \mu e^{|a|_{\oplus} s} && \text{if } a \in \mathbb{S}^{\bullet}, \end{aligned}$$

where $a \in \mathbb{S}$, $\mu \in \mathbb{R}_0$, and $s \in \mathbb{R}_0^+$.

In the remainder of this paper the first two arguments of $\mathcal{F}$ will most of the time be fixed and we shall only consider $\mathcal{F}$ in function of the third argument, i.e., for a given $a \in \mathbb{S}$ and $\mu \in \mathbb{R}_0$ we consider the function $\mathcal{F}(a, \mu, \cdot)$. Note that if $x \in \mathbb{R}_\varepsilon$ and $\mu \in \mathbb{R}_0$, then we have

$$\begin{aligned} \mathcal{F}(x, \mu, s) &= |\mu| e^{xs}, \\ \mathcal{F}(\ominus x, \mu, s) &= -|\mu| e^{xs}, \\ \mathcal{F}(x^{\bullet}, \mu, s) &= \mu e^{xs} \end{aligned}$$

for all $s \in \mathbb{R}_0^+$. Furthermore, $\mathcal{F}(\varepsilon, \mu, \cdot) = 0$ for all $\mu \in \mathbb{R}_0$ since we have $e^{\varepsilon s} = 0$ for all $s \in \mathbb{R}_0^+$ by definition.

For a given $\mu \in \mathbb{R}_0$ the number $a \in \mathbb{S}$ will be mapped by $\mathcal{F}$ to an exponential function $s \mapsto \nu e^{|a|_\oplus s}$, where $\nu = |\mu|$, $\nu = -|\mu|$, or $\nu = \mu$ depending on the max-algebraic sign of a. In order to reverse this process, we define the mapping $\mathcal{R}$, which we shall call the *reverse mapping* and which will map a function that is asymptotically equivalent to an exponential function $s \mapsto \nu e^{|a|_\oplus s}$ in the neighborhood of ∞ to the number $|a|_\oplus$ or $\ominus |a|_\oplus$ depending on the sign of ν. More specifically, if f is a real function, if $x \in \mathbb{R}_\varepsilon$, and if $\mu \in \mathbb{R}_0$, then we have

$$\begin{aligned} f(s) \ &\sim \ |\mu| e^{xs} \ , \ s \to \infty \quad &\Rightarrow \quad \mathcal{R}(f) = x, \\ f(s) \ &\sim \ -|\mu| e^{xs} \ , \ s \to \infty \quad &\Rightarrow \quad \mathcal{R}(f) = \ominus x \ . \end{aligned}$$

Note that $\mathcal{R}$ will always map a function that is asymptotically equivalent to an exponential function in the neighborhood of ∞ to a signed number and never to a balanced number that is different from ε. Furthermore, for a fixed $\mu \in \mathbb{R}_0$ the mappings $a \mapsto \mathcal{F}(a, \mu, \cdot)$ and $\mathcal{R}$ are not each other's inverse since these mappings are not bijections as shown by the following example.

Example 4.1. Let $\mu = 1$. We have $\mathcal{F}(2, \mu, s) = e^{2s}$ and $\mathcal{F}(2^\bullet, \mu, s) = e^{2s}$ for all $s \in \mathbb{R}_0^+$. So $\mathcal{R}(\mathcal{F}(2^\bullet, \mu, \cdot)) = 2 \neq 2^\bullet$.

Consider the real functions f and g defined by $f(s) = e^{2s}$ and $g(s) = e^{2s} + 1$. We have $f(s) \sim g(s) \sim e^{2s}$, $s \to \infty$. Hence, $\mathcal{R}(f) = \mathcal{R}(g) = 2$. So $\mathcal{F}(\mathcal{R}(g), \mu, \cdot) = f \neq g$. □

Let $\mu \in \mathbb{R}_0$. It is easy to verify that, in general, we have $\mathcal{R}(\mathcal{F}(a, \mu, \cdot)) = a$ if $a \in \mathbb{S}^\oplus \cup \mathbb{S}^\ominus$, $\mathcal{R}(\mathcal{F}(a, \mu, \cdot)) = |a|_\oplus$ if $a \in \mathbb{S}^\bullet$ and $\mu > 0$, and $\mathcal{R}(\mathcal{F}(a, \mu, \cdot)) = \ominus |a|_\oplus$ if $a \in \mathbb{S}^\bullet$ and $\mu < 0$. Furthermore, if f is a real function that is asymptotically equivalent to an exponential function in the neighborhood of ∞, then we have $\mathcal{F}(\mathcal{R}(f), \mu, s) \sim f(s)$, $s \to \infty$.

For all $a, b, c \in \mathbb{S}$ we have

$$a \oplus b = c \ \Rightarrow \ \left\{ \begin{array}{l} \exists \mu_a, \mu_b, \mu_c \in \mathbb{R}_0 \text{ such that} \\ \mathcal{F}(a, \mu_a, s) + \mathcal{F}(b, \mu_b, s) \ \sim \ \mathcal{F}(c, \mu_c, s) \ , \ s \to \infty, \end{array} \right. \tag{16}$$

$$\left. \begin{array}{l} \exists \mu_a, \mu_b, \mu_c \in \mathbb{R}_0 \text{ such that} \\ \mathcal{F}(a, \mu_a, s) + \mathcal{F}(b, \mu_b, s) \ \sim \ \mathcal{F}(c, \mu_c, s) \ , \ s \to \infty \end{array} \right\} \ \Rightarrow \ a \oplus b \ \nabla \ c, \tag{17}$$

$$a \otimes b = c \ \Rightarrow \ \left\{ \begin{array}{l} \exists \mu_a, \mu_b, \mu_c \in \mathbb{R}_0 \text{ such that} \\ \mathcal{F}(a, \mu_a, s) \cdot \mathcal{F}(b, \mu_b, s) = \mathcal{F}(c, \mu_c, s) \quad \text{for all } s \in \mathbb{R}_0^+, \end{array} \right. \tag{18}$$

$$\left. \begin{array}{l} \exists \mu_a, \mu_b, \mu_c \in \mathbb{R}_0 \text{ such that} \\ \mathcal{F}(a, \mu_a, s) \cdot \mathcal{F}(b, \mu_b, s) = \mathcal{F}(c, \mu_c, s) \quad \text{for all } s \in \mathbb{R}_0^+ \end{array} \right\} \ \Rightarrow \ a \otimes b \ \nabla \ c. \tag{19}$$

As a consequence, we could say that the mapping $\mathcal{F}$ provides a link between the structure $(\mathcal{R}_\mathrm{e}^+, +, \times)$ and $\mathbb{R}_\mathrm{max} = (\mathbb{R}_\varepsilon, \oplus, \otimes)$, and a link between the structure $(\mathcal{R}_\mathrm{e}, +, \times)$ and $\mathbb{S}_\mathrm{max} = (\mathbb{S}, \oplus, \otimes)$.

Remark 4.2. The balance in (17) results from the fact that we can have cancellation of equal terms with opposite sign in $(\mathcal{R}_\mathrm{e}^+, +, \times)$, whereas this is, in general, not possible in the symmetrized max-plus algebra since $\forall a \in \mathbb{S} \setminus \{\varepsilon\} : a \ominus a \neq \varepsilon$.

The following example shows that the balance on the right-hand side of (19) is also necessary: we have $\mathcal{F}(0, 1, s) \cdot \mathcal{F}(0, 1, s) = 1 \cdot 1 = 1 = \mathcal{F}(0^\bullet, 1, s)$ for all $s \in \mathbb{R}_0^+$, but $0 \otimes 0 = 0 \neq 0^\bullet$.

We have $1 \oplus 2 = 2 \nabla 3^\bullet$. However, there do not exist real numbers μ_1, μ_2, $\mu_3 \in \mathbb{R}_0$ such that

$$\mathcal{F}(1, \mu_1, s) + \mathcal{F}(2, \mu_2, s) \;\sim\; \mathcal{F}(3^\bullet, \mu_3, s)\ ,\ s \to \infty$$

or equivalently

$$|\mu_1|\, e^s \;+\; |\mu_2|\, e^{2s} \;\sim\; \mu_3\, e^{3s}\ ,\ s \to \infty.$$

This implies that, in general, (16) does not hold any more if we replace the equality on the left-hand side by a balance.

In a similar way we can show that, in general, $a \otimes b \nabla c$ with $a, b, c, \in \mathbb{S}$ does not imply that there exist real numbers μ_a, μ_b, $\mu_c \in \mathbb{R}_0$ such that $\mathcal{F}(a, \mu_a, s) \cdot \mathcal{F}(b, \mu_b, s) = \mathcal{F}(c, \mu_c, s)$ for all $s \in \mathbb{R}_0^+$. □

We extend the mapping $\mathcal{F}$ to matrices as follows. If $A \in \mathbb{S}^{m \times n}$ and if $M \in \mathbb{R}_0^{m \times n}$, then $\tilde{A} = \mathcal{F}(A, M, \cdot)$ is a real $m \times n$ matrix-valued function with domain of definition $\mathbb{R}_0^+$ and with $\tilde{a}_{ij}(s) = \mathcal{F}(a_{ij}, m_{ij}, s)$ for all i, j. Note that the mapping is performed entrywise. The reverse mapping $\mathcal{R}$ is extended to matrices in a similar way: if $\tilde{A}$ is a real matrix-valued function with entries that are asymptotically equivalent to an exponential in the neighborhood of ∞, then $(\mathcal{R}(\tilde{A}))_{ij} = \mathcal{R}(\tilde{a}_{ij})$ for all i, j. If A, B, and C are matrices with entries in $\mathbb{S}$, we have

$$A \oplus B = C \;\Rightarrow\; \left\{ \begin{array}{l} \exists M_A, M_B, M_C \text{ such that} \\ \mathcal{F}(A, M_A, s) + \mathcal{F}(B, M_B, s) \;\sim\; \mathcal{F}(C, M_C, s)\ ,\ s \to \infty, \end{array} \right. \tag{20}$$

$$\left. \begin{array}{l} \exists M_A, M_B, M_C \text{ such that} \\ \mathcal{F}(A, M_A, s) + \mathcal{F}(B, M_B, s) \;\sim\; \mathcal{F}(C, M_C, s)\ ,\ s \to \infty \end{array} \right\} \;\Rightarrow\; A \oplus B \;\nabla\; C, \tag{21}$$

$$A \otimes B = C \;\Rightarrow\; \left\{ \begin{array}{l} \exists M_A, M_B, M_C \text{ such that} \\ \mathcal{F}(A, M_A, s) \cdot \mathcal{F}(B, M_B, s) \;\sim\; \mathcal{F}(C, M_C, s)\ ,\ s \to \infty, \end{array} \right. \tag{22}$$

$$\left. \begin{array}{l} \exists M_A, M_B, M_C \text{ such that} \\ \mathcal{F}(A, M_A, s) \cdot \mathcal{F}(B, M_B, s) \;\sim\; \mathcal{F}(C, M_C, s)\ ,\ s \to \infty \end{array} \right\} \;\Rightarrow\; A \otimes B \;\nabla\; C. \tag{23}$$

Example 4.3. Let $A = \left[\begin{smallmatrix} 0 & \varepsilon \\ \ominus 1 & \ominus 2 \end{smallmatrix} \right]$ and $B = \left[\begin{smallmatrix} -3 & 1 \\ 2^\bullet & \ominus 0 \end{smallmatrix} \right]$. Hence, $A \otimes B = \left[\begin{smallmatrix} -3 & 1 \\ 4^\bullet & 2^\bullet \end{smallmatrix} \right]$. Let M, N, and $P \in \mathbb{R}_0^{2 \times 2}$. In general we have

$$\mathcal{F}(A, M, s) = \begin{bmatrix} |m_{11}| & 0 \\ -|m_{21}|\, e^s & -|m_{22}|\, e^{2s} \end{bmatrix},$$

$$\mathcal{F}(B, N, s) = \begin{bmatrix} |n_{11}|\, e^{-3s} & |n_{12}|\, e^s \\ n_{21}\, e^{2s} & -|n_{22}| \end{bmatrix},$$

$$\mathcal{F}(A \otimes B, P, s) = \begin{bmatrix} |p_{11}|\, e^{-3s} & |p_{12}|\, e^s \\ p_{21}\, e^{4s} & p_{22}\, e^{2s} \end{bmatrix}$$

for all $s \in \mathbb{R}_0^+$. Furthermore,

$$\begin{aligned} &\mathcal{F}(A, M, s) \cdot \mathcal{F}(B, N, s) \\ &\quad = \begin{bmatrix} |m_{11}|\, |n_{11}|\, e^{-3s} & |m_{11}|\, |n_{12}|\, e^s \\ -|m_{21}|\, |n_{11}|\, e^{-2s} \;-\; |m_{22}|\, n_{21}\, e^{4s} & (-|m_{21}|\, |n_{12}| \;+\; |m_{22}|\, |n_{22}|)\, e^{2s} \end{bmatrix} \end{aligned}$$

for all $s \in \mathbb{R}_0^+$.

If $-|m_{21}||n_{12}| + |m_{22}||n_{22}| \neq 0$ and if we take

$$p_{11} = |m_{11}|\,|n_{11}|\,,\;\; p_{12} = |m_{11}|\,|n_{12}|\,,$$
$$p_{21} = -|m_{22}|\,n_{21}\,,\;\; p_{22} = -|m_{21}|\,|n_{12}| + |m_{22}|\,|n_{22}|\,,$$

then we have $\mathcal{F}(A,M,s)\cdot\mathcal{F}(B,N,s) \sim \mathcal{F}(A\otimes B,P,s)\,,\; s\to\infty$.

If we take $m_{ij} = n_{ij} = 1$ for all i,j, we get

$$\mathcal{F}(A,s)\cdot\mathcal{F}(B,s) \sim \begin{bmatrix} e^{-3s} & e^{s} \\ -e^{4s} & 0 \end{bmatrix} \stackrel{\text{def}}{=} \tilde{C}(s)\,,\; s\to\infty.$$

The reverse mapping results in $C = \mathcal{R}(\tilde{C}) = \left[\begin{smallmatrix} -3 & 1 \\ \ominus 4 & \varepsilon \end{smallmatrix}\right]$. Note that $A\otimes B \nabla C$.

Taking $m_{ij} = n_{ij} = (-1)^{(i+j)}(i+j)$ for all i,j leads to

$$\mathcal{F}(A,s)\cdot\mathcal{F}(B,s) \sim \begin{bmatrix} 4e^{-3s} & 6e^{s} \\ 12e^{4s} & 7e^{2s} \end{bmatrix} \stackrel{\text{def}}{=} \tilde{D}(s)\,,\; s\to\infty.$$

The reverse mapping results in $D = \mathcal{R}(\tilde{D}) = \left[\begin{smallmatrix} -3 & 1 \\ 4 & 2 \end{smallmatrix}\right]$ and again we have $A\otimes B \nabla D$. ∎

5. The QR decomposition and the SVD in the symmetrized max-plus algebra. In [10] we have used the mapping from $\mathbb{S}_{\max}$ to $(\mathcal{R}_{\mathrm{e}},+,\times)$ and the reverse mapping $\mathcal{R}$ to prove the existence of a kind of SVD in $\mathbb{S}_{\max}$. The proof of [10] is based on the *analytic SVD*. In this section we present an alternative proof for the existence theorem of the max-algebraic SVD. The major advantage of the new proof technique that will be developed in this section over the one of [10] is that it can easily be extended to prove the existence of many other matrix decompositions in the symmetrized max-plus algebra such as the max-algebraic QR decomposition, the max-algebraic LU decomposition, the max-algebraic eigenvalue decomposition (for symmetric matrices), and so on. This proof technique consists of transforming a matrix with entries in $\mathbb{S}$ to a matrix-valued function with exponential entries (using the mapping $\mathcal{F}$), applying an algorithm from linear algebra, and transforming the result back to the symmetrized max-plus algebra (using the mapping $\mathcal{R}$).

5.1. Sums and series of exponentials. The entries of the matrices that are used in the existence proofs for the max-algebraic QR decomposition and the max-algebraic SVD that will be presented in this section are sums or series of exponentials. Therefore, we first study some properties of this kind of functions.

DEFINITION 5.1 (sum or series of exponentials). *Let $\mathcal{S}_{\mathrm{e}}$ be the set of real functions that are analytic and that can be written as a (possibly infinite but absolutely convergent) sum of exponentials in a neighborhood of ∞:*

$$\mathcal{S}_{\mathrm{e}} = \Big\{ f : A\to\mathbb{R} \;\Big|\; A\subseteq\mathbb{R},\; \exists K\in\mathbb{R}_0^+ \text{ such that } [K,\infty)\subseteq A \text{ and}$$
$$f \text{ is analytic in } [K,\infty), \text{ and either}$$
$$\forall x\geq K : f(x) = \sum_{i=0}^{n} \alpha_i e^{a_i x} \tag{24}$$
$$\text{with } n\in\mathbb{N},\; \alpha_i\in\mathbb{R}_0,\; a_i\in\mathbb{R}_\varepsilon \text{ for all } i \text{ and } a_0 > a_1 > \cdots > a_n, \text{ or}$$
$$\forall x\geq K : f(x) = \sum_{i=0}^{\infty} \alpha_i e^{a_i x} \tag{25}$$

$$\text{with } \alpha_i \in \mathbb{R}_0,\ a_i \in \mathbb{R},\ a_i > a_{i+1} \text{ for all } i\,, \lim_{i\to\infty} a_i = \varepsilon \text{ and}$$
$$\text{where the series converges absolutely for every } x \geq K \Big\}.$$

If $f \in \mathcal{S}_{\mathrm{e}}$, then the largest exponent in the sum or the series of exponentials that corresponds to f is called the *dominant exponent* of f.

Recall that by definition we have $e^{\varepsilon s} = 0$ for all $s \in \mathbb{R}_0^+$. Since we allow exponents that are equal to $\varepsilon = -\infty$ in the definition of $\mathcal{S}_{\mathrm{e}}$, the zero function also belongs to $\mathcal{S}_{\mathrm{e}}$. Since we require that the sequence of the exponents that appear in (24) or (25) is decreasing and since the coefficients cannot be equal to zero, any sum of exponentials of the form (24) or (25) that corresponds to the zero function consists of exactly one term, e.g., $1 \cdot e^{\varepsilon x}$.

If $f \in \mathcal{S}_{\mathrm{e}}$ is a series of the form (25), then the set $\{a_i \,|\, i = 0, 1, \ldots, \infty\}$ has no finite accumulation point since the sequence $\{a_i\}_{i=0}^{\infty}$ is decreasing and unbounded from below. Note that series of the form (25) are related to (generalized) Dirichlet series [23].

The behavior of the functions in $\mathcal{S}_{\mathrm{e}}$ in the neighborhood of ∞ is given by the following property.

LEMMA 5.2. *Every function $f \in \mathcal{S}_{\mathrm{e}}$ is asymptotically equivalent to an exponential in the neighborhood of ∞:*

$$f \in \mathcal{S}_{\mathrm{e}} \ \Rightarrow\ f(x) \ \sim\ \alpha_0 e^{a_0 x}\,,\ x \to \infty$$

for some $\alpha_0 \in \mathbb{R}_0$ and some $a_0 \in \mathbb{R}_\varepsilon$.

Proof. See Appendix A. ∎

The set $\mathcal{S}_{\mathrm{e}}$ is closed under elementary operations such as additions, multiplications, subtractions, divisions, square roots, and absolute values.

PROPOSITION 5.3. *If f and g belong to $\mathcal{S}_{\mathrm{e}}$, then ρf, $f+g$, $f-g$, fg, f^l, and $|f|$ also belong to $\mathcal{S}_{\mathrm{e}}$ for any $\rho \in \mathbb{R}$ and any $l \in \mathbb{N}$.*

Furthermore, if there exists a real number P such that $f(x) \neq 0$ for all $x \geq P$, then the functions $\frac{1}{f}$ and $\frac{g}{f}$ restricted to $[P, \infty)$ also belong to $\mathcal{S}_{\mathrm{e}}$.

If there exists a real number Q such that $f(x) > 0$ for all $x \geq Q$, then the function $\sqrt{f}$ restricted to $[Q, \infty)$ also belongs to $\mathcal{S}_{\mathrm{e}}$.

Proof. See Appendix B. ∎

5.2. The max-algebraic QR decomposition. Let $\tilde{A}$ and $\tilde{R}$ be real $m \times n$ matrix-valued functions and let $\tilde{Q}$ be a real $m \times m$ matrix-valued function. Suppose that these matrix-valued functions are defined in $J \subseteq \mathbb{R}$. If $\tilde{Q}(s)\,\tilde{R}(s) = \tilde{A}(s)$, $\tilde{Q}^T(s)\,\tilde{Q}(s) = I_m$, and $\tilde{R}(s)$ is an upper triangular matrix for all $s \in J$, then we say that $\tilde{Q}\tilde{R}$ is a *path of QR decompositions* of $\tilde{A}$ on J. A path of SVDs is defined in a similar way.

Note that if $\tilde{Q}\tilde{R}$ is a path of QR decompositions of $\tilde{A}$ on J, then we have $\|\tilde{R}(s)\|_{\mathrm{F}} = \|\tilde{A}(s)\|_{\mathrm{F}}$ for all $s \in J$. Now we prove that for a matrix with entries in $\mathcal{S}_{\mathrm{e}}$ there exists a path of QR decompositions with entries that also belong to $\mathcal{S}_{\mathrm{e}}$. Next we use this result to prove the existence of a max-algebraic analogue of the QR decomposition.

PROPOSITION 5.4. *If $\tilde{A} \in \mathcal{S}_{\mathrm{e}}^{m\times n}$, then there exists a path of QR decompositions $\tilde{Q}\tilde{R}$ of $\tilde{A}$ for which the entries of $\tilde{Q}$ and $\tilde{R}$ belong to $\mathcal{S}_{\mathrm{e}}$.*

Proof. To compute the QR decomposition of a matrix with real entries we can use the Givens QR algorithm (see [14]). The operations used in this algorithm are additions, multiplications, subtractions, divisions, and square roots. Furthermore, the number of operations used in this algorithm is finite.

So if we apply this algorithm to a matrix-valued function $\tilde{A}$ with entries in $\mathcal{S}_\mathrm{e}$, then the entries of the resulting matrix-valued functions $\tilde{Q}$ and $\tilde{R}$ will also belong to $\mathcal{S}_\mathrm{e}$ by Proposition 5.3. □

THEOREM 5.5 (max-algebraic QR decomposition). *If $A \in \mathbb{S}^{m\times n}$, then there exist a matrix $Q \in (\mathbb{S}^\vee)^{m\times m}$ and a max-algebraic upper triangular matrix $R \in (\mathbb{S}^\vee)^{m\times n}$ such that*

$$A \nabla Q \otimes R \tag{26}$$

with $Q^T \otimes Q \nabla E_m$ and $\|R\|_\oplus = \|A\|_\oplus$.

Every decomposition of the form (26) *that satisfies the above conditions is called a max-algebraic QR decomposition of A.*

Proof. If $A \in \mathbb{S}^{m\times n}$ has entries that are not signed, we can always define a matrix $\hat{A} \in (\mathbb{S}^\vee)^{m\times n}$ such that

$$\hat{a}_{ij} = \begin{cases} a_{ij} & \text{if } a_{ij} \text{ is signed,} \\ |a_{ij}|_\oplus & \text{if } a_{ij} \text{ is not signed} \end{cases}$$

for all i, j. Since $|\hat{a}_{ij}|_\oplus = |a_{ij}|_\oplus$ for all i, j, we have $\|\hat{A}\|_\oplus = \|A\|_\oplus$. Moreover, we have

$$\forall a, b \in \mathbb{S} : \; a \nabla b \;\Rightarrow\; a^\bullet \nabla b,$$

which means that if $\hat{A} \nabla Q \otimes R$, then also $A \nabla Q \otimes R$. Therefore, it is sufficient to prove this theorem for signed matrices A.

So from now on we assume that A is signed. We construct $\tilde{A} = \mathcal{F}(A, M, \cdot)$, where $M \in \mathbb{R}^{m\times n}$ with $m_{ij} = 1$ for all i, j. Hence, $\tilde{a}_{ij}(s) = \gamma_{ij} e^{c_{ij} s}$ for all $s \in \mathbb{R}_0^+$ and for all i, j with $\gamma_{ij} \in \{-1, 1\}$ and $c_{ij} = |a_{ij}|_\oplus \in \mathbb{R}_\varepsilon$ for all i, j. Note that the entries of $\tilde{A}$ belong to $\mathcal{S}_\mathrm{e}$. By Proposition 5.4 there exists a path of QR decompositions of $\tilde{A}$. So there exists a positive real number L and matrix-valued functions $\tilde{Q}$ and $\tilde{R}$ with entries in $\mathcal{S}_\mathrm{e}$ such that

$$\tilde{A}(s) = \tilde{Q}(s)\, \tilde{R}(s) \qquad \text{for all } s \geq L, \tag{27}$$

$$\tilde{Q}^T(s)\, \tilde{Q}(s) = I_m \qquad \text{for all } s \geq L, \tag{28}$$

$$\|\tilde{R}(s)\|_\mathrm{F} = \|\tilde{A}(s)\|_\mathrm{F} \qquad \text{for all } s \geq L. \tag{29}$$

The entries of $\tilde{Q}$ and $\tilde{R}$ belong to $\mathcal{S}_\mathrm{e}$ and are thus asymptotically equivalent to an exponential in the neighborhood of ∞ by Lemma 5.2.

If we define $Q = \mathcal{R}(\tilde{Q})$ and $R = \mathcal{R}(\tilde{R})$, then Q and R have signed entries. If we apply the reverse mapping $\mathcal{R}$ to (27)–(29), we get

$$A \nabla Q \otimes R\,, \quad Q^T \otimes Q \nabla E_m, \;\text{ and }\; \|R\|_\oplus = \|A\|_\oplus\,. \qquad \square$$

If f, g, and h belong to $\mathcal{S}_\mathrm{e}$, then they are asymptotically equivalent to an exponential in the neighborhood of ∞ by Lemma 5.2. So if L is large enough, then $f(L) \geq 0$ and $g(L) \geq h(L)$ imply that $f(s) \geq 0$ and $g(s) \geq h(s)$ for all $s \in [L, \infty)$. This fact and the fact that $\mathcal{S}_\mathrm{e}$ is closed under some elementary algebraic operations explain why many algorithms from linear algebra—such as the Givens QR algorithm and Kogbetliantz's SVD algorithm (see section 5.3)—also work for matrices with entries that belong to $\mathcal{S}_\mathrm{e}$ instead of $\mathbb{R}$. If we apply an algorithm from linear algebra to a matrix-valued function $\tilde{A}$ with entries in $\mathcal{S}_\mathrm{e}$ that is defined on some interval $[L, \infty)$, we are in fact

applying this algorithm on the (constant) matrix $\tilde{A}(s)$ for every value of $s \in [L, \infty)$ in parallel.

If QR is a QR decomposition of a matrix $A \in \mathbb{R}^{m\times n}$, then we always have $\|R\|_{\mathrm{F}} = \|A\|_{\mathrm{F}}$ since Q is an orthogonal matrix. However, the following example shows that $A \nabla Q \otimes R$ and $Q^T \otimes Q \nabla E_m$ do not always imply that $\|R\|_{\oplus} = \|A\|_{\oplus}$.

Example 5.6. Consider

$$A = \begin{bmatrix} \ominus 0 & 0 & 0 \\ 0 & \ominus 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Without the condition $\|R\|_{\oplus} = \|A\|_{\oplus}$ every max-algebraic product of the form

$$Q \otimes R(\rho) = \begin{bmatrix} \ominus 0 & 0 & 0 \\ 0 & \ominus 0 & 0 \\ 0 & 0 & \ominus 0 \end{bmatrix} \otimes \begin{bmatrix} 0 & \varepsilon & \rho \\ \varepsilon & 0 & \rho \\ \varepsilon & \varepsilon & \rho \end{bmatrix} = \begin{bmatrix} \ominus 0 & 0 & \rho^\bullet \\ 0 & \ominus 0 & \rho^\bullet \\ 0 & 0 & \rho^\bullet \end{bmatrix}$$

with $\rho \geq 0$ would have been a max-algebraic QR decomposition of A. However, since $\|R(\rho)\|_{\oplus} = \rho$ if $\rho \geq 0$ and since $\|A\|_{\oplus} = 0$, we do not have $\|R\|_{\oplus} = \|A\|_{\oplus}$ if $\rho > 0$. □

This example explains why we have included the condition $\|R\|_{\oplus} = \|A\|_{\oplus}$ in the definition of the max-algebraic QR decomposition.

Now we explain why we really need the symmetrized max-plus algebra $\mathbb{S}_{\max}$ to define the max-algebraic QR decomposition: we shall show that the class of matrices with entries in $\mathbb{R}_\varepsilon$ that have max-algebraic QR decompositions for which the entries of Q and R belong to $\mathbb{R}_\varepsilon$ is rather limited. Let $A \in \mathbb{R}_\varepsilon^{m\times n}$ and let $Q \otimes R$ be a max-algebraic QR decomposition of A for which the entries of Q and R belong to $\mathbb{R}_\varepsilon$. Since the entries of A, Q, and R are signed, it follows from Proposition 3.6 that the balances $A \nabla Q \otimes R$ and $Q^T \otimes Q \nabla E_m$ result in $A = Q \otimes R$ and $Q^T \otimes Q = E_m$. It is easy to verify that we can only have $Q^T \otimes Q = E_m$ if every column and every row of Q contains exactly one entry that is equal to zero and if all the other entries of Q are equal to ε. Hence, Q is max-algebraic permutation matrix. As a consequence, A has to be a row-permuted max-algebraic upper triangular matrix.

So only row-permuted max-algebraic upper triangular matrices with entries in $\mathbb{R}_\varepsilon$ have a max-algebraic QR decomposition with entries in $\mathbb{R}_\varepsilon$. This could be compared with the class of real matrices in linear algebra that have a QR decomposition with only nonnegative entries: using an analogous reasoning one can prove that this class coincides with the set of the real row-permuted upper triangular matrices. Furthermore, it is obvious that every QR decomposition in $\mathbb{R}_{\max}$ is also a QR decomposition in $\mathbb{S}_{\max}$.

5.3. The max-algebraic SVD. Now we give an alternative proof for the existence theorem of the max-algebraic SVD. In this proof we shall use Kogbetliantz's SVD algorithm [20], which can be considered an extension of Jacobi's method for the computation of the eigenvalue decomposition of a real symmetric matrix. We now state the main properties of this algorithm. The explanation below is mainly based on [4] and [17].

A Givens matrix is a square matrix of the form

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & \cos(\theta) & \cdots & \sin(\theta) & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \ddots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & -\sin(\theta) & \cdots & \cos(\theta) & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

The off-norm of the matrix $A \in \mathbb{R}^{m\times n}$ is defined by

$$\|A\|_{\text{off}} = \sqrt{\sum_{i=1}^{n} \sum_{j=1,\, j\neq i}^{n} a_{ij}^2},$$

where the empty sum is equal to zero by definition (so if A is a 1×1 matrix, then we have $\|A\|_{\text{off}} = 0$). Let $A \in \mathbb{R}^{m\times n}$. Since USV^T is an SVD of A if and only if VS^TU^T is an SVD of A^T, we may assume without loss of generality that $m \geq n$. Before applying Kogbetliantz's SVD algorithm we compute a QR decomposition of A:

$$A = Q \begin{bmatrix} R \\ O_{(m-n)\times n} \end{bmatrix},$$

where R is an $n \times n$ upper triangular matrix.

Now we apply Kogbetliantz's SVD algorithm to R. In this algorithm a sequence of matrices is generated as follows:

$$\begin{aligned} &U_0 = I_n\,, \quad V_0 = I_n\,, \quad S_0 = R\,, \\ &U_k = U_{k-1}G_k\,, \quad V_k = V_{k-1}H_k\,, \quad S_k = G_k^T S_{k-1} H_k \qquad \text{for } k = 1, 2, 3, \ldots \end{aligned}$$

such that $\|S_k\|_{\text{off}}$ decreases monotonously as k increases. So S_k tends more and more to a diagonal matrix as the iteration process progresses. The absolute values of the diagonal entries of S_k will converge to the singular values of R as k goes to ∞.

The matrices G_k and H_k are Givens matrices that are chosen such that $(S_k)_{i_k j_k} = (S_k)_{j_k i_k} = 0$ for some ordered pair of indices (i_k, j_k). As a result we have

$$\|S_k\|_{\text{off}}^2 = \|S_{k-1}\|_{\text{off}}^2 - (S_{k-1})_{i_k j_k}^2 - (S_{k-1})_{j_k i_k}^2 \ .$$

Since the matrices G_k and H_k are orthogonal for all $k \in \mathbb{N}_0$, we have

$$\|S_k\|_{\text{F}} = \|R\|_{\text{F}}\,, \quad R = U_k S_k V_k^T, \quad U_k^T U_k = I_n, \text{ and } V_k^T V_k = I_n \tag{30}$$

for all $k \in \mathbb{N}$.

We shall use the row-cyclic version of Kogbetliantz's SVD algorithm: in each cycle the indices i_k and j_k are chosen such that the entries in the strictly upper triangular part of the S_k's are selected row by row. This yields the following sequence for the ordered pairs of indices (i_k, j_k):

$$(1,2) \to (1,3) \to \cdots \to (1,n) \to (2,3) \to (2,4) \to \cdots \to (n-1,n).$$

A full cycle $(1,2) \to \cdots \to (n-1,n)$ is called a *sweep*. Note that a sweep corresponds to $N = \frac{(n-1)n}{2}$ iterations. Sweeps are repeated until S_k becomes diagonal. If we have an upper triangular matrix at the beginning of a sweep, then we shall have a lower triangular matrix after the sweep and vice versa.

For triangular matrices the row-cyclic Kogbetliantz algorithm is globally convergent [11, 17]. Furthermore, for triangular matrices the convergence of this algorithm is quadratic if k is large enough [2, 3, 15, 16, 31]:

$$\exists K \in \mathbb{N} \text{ such that } \forall k \geq K : \|S_{k+N}\|_{\text{off}} \leq \gamma \|S_k\|_{\text{off}}^2 \tag{31}$$

for some constant γ that does not depend on k, under the assumption that diagonal entries that correspond to the same singular value or that are affiliated with the same cluster of singular values occupy successive positions on the diagonal. This assumption is not restrictive since we can always reorder the diagonal entries of S_k by inserting an extra step in which we select a permutation matrix $\hat{P} \in \mathbb{R}^{n\times n}$ such that the diagonal entries of $S_{k+1} = \hat{P}^T S_k \hat{P}$ exhibit the required ordering. Note that $\|S_{k+1}\|_{\text{F}} = \|S_k\|_{\text{F}}$. If we define $U_{k+1} = U_k \hat{P}$ and $V_{k+1} = V_k \hat{P}$, then U_{k+1} and V_{k+1} are orthogonal since $\hat{P}^T \hat{P} = I_n$. We also have

$$U_{k+1} S_{k+1} V_{k+1}^T = \left(U_k \hat{P}\right) \left(\hat{P}^T S_k \hat{P}\right) \left(\hat{P}^T V_k^T\right) = U_k S_k V_k^T = R.$$

Furthermore, once the diagonal entries have the required ordering, they hold it provided that k is sufficiently large [15].

If we define $S = \lim_{k\to\infty} S_k$, $U = \lim_{k\to\infty} U_k$, and $V = \lim_{k\to\infty} V_k$, then S is a diagonal matrix, U and V are orthogonal matrices, and $USV^T = R$. We make all the diagonal entries of S nonnegative by multiplying S with an appropriate diagonal matrix D. Next we construct a permutation matrix P such that the diagonal entries of $P^T SDP$ are arranged in descending order. If we define $U_R = UP$, $S_R = P^T SDP$, and $V_R = VD^{-1}P$, then U_R and V_R are orthogonal, the diagonal entries of S_R are nonnegative and ordered, and

$$U_R S_R V_R^T = (UP)\left(P^T SDP\right)\left(P^T D^{-1} V^T\right) = USV^T = R.$$

Hence, $U_R S_R V_R^T$ is an SVD of R. If we define

$$U_A = Q \begin{bmatrix} U_R & O_{n\times(m-n)} \\ O_{(m-n)\times n} & I_{m-n} \end{bmatrix}, \; S_A = \begin{bmatrix} S_R \\ O_{(m-n)\times n} \end{bmatrix}, \text{ and } \; V_A = V_R,$$

then $U_A S_A V_A^T$ is an SVD of A.

THEOREM 5.7 (max-algebraic SVD). *Let $A \in \mathbb{S}^{m\times n}$ and let $r = \min(m,n)$. Then there exist a max-algebraic diagonal matrix $\Sigma \in \mathbb{R}_\varepsilon^{m\times n}$ and matrices $U \in (\mathbb{S}^\vee)^{m\times m}$ and $V \in (\mathbb{S}^\vee)^{n\times n}$ such that*

$$A \nabla U \otimes \Sigma \otimes V^T \tag{32}$$

with $U^T \otimes U \nabla E_m$, $V^T \otimes V \nabla E_n$, and $\|A\|_\oplus = \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$, where $\sigma_i = (\Sigma)_{ii}$ for $i = 1, 2, \ldots, r$.

Every decomposition of the form (32) *that satisfies the above conditions is called a max-algebraic SVD of A.*

Proof. Using a reasoning that is similar to the one that has been used at the beginning of the proof of Theorem 5.5, we can show that it is sufficient to prove this theorem for signed matrices A. So from now on we assume that A is signed.

Define $c = \|A\|_{\oplus}$. If $c = \varepsilon$, then $A = \mathcal{E}_{m\times n}$. If we take $U = E_m$, $\Sigma = \mathcal{E}_{m\times n}$ and $V = E_n$, we have $A = U \otimes \Sigma \otimes V^T$, $U^T \otimes U = E_m$, $V^T \otimes V = E_n$, and $\sigma_1 = \sigma_2 = \cdots = \sigma_r = \varepsilon = \|A\|_{\oplus}$. So $U \otimes \Sigma \otimes V^T$ is a max-algebraic SVD of A.

From now on we assume that $c \neq \varepsilon$. We may assume without loss of generality that $m \geq n$: if $m < n$, we can apply the subsequent reasoning to A^T since $A \nabla U \otimes \Sigma \otimes V^T$ if and only if $A^T \nabla V \otimes \Sigma^T \otimes U^T$. So $U \otimes \Sigma \otimes V^T$ is a max-algebraic SVD of A if and only if $V \otimes \Sigma^T \otimes U^T$ is a max-algebraic SVD of A^T.

Now we distinguish between two different situations depending on whether or not all the a_{ij}'s have a finite max-absolute value. In Remark 5.8 we shall explain why this distinction is necessary.

Case 1. All the a_{ij}'s have a finite max-absolute value.

We construct $\tilde{A} = \mathcal{F}(A, M, \cdot)$, where $M \in \mathbb{R}^{m\times n}$ with $m_{ij} = 1$ for all i, j. The entries of $\tilde{A}$ belong to $\mathcal{S}_{\mathrm{e}}$.

In order to determine a path of SVDs of $\tilde{A}$, we first compute a path of QR decompositions of $\tilde{A}$ on $\mathbb{R}_0^+$:

$$\tilde{A} = \tilde{Q} \begin{bmatrix} \tilde{R} \\ O_{(m-n)\times n} \end{bmatrix},$$

where $\tilde{R}$ is an $n \times n$ upper triangular matrix-valued function. By Proposition 5.4 the entries of $\tilde{Q}$ and $\tilde{R}$ belong to $\mathcal{S}_{\mathrm{e}}$.

Now we use the row-cyclic Kogbetliantz algorithm to compute a path of SVDs of $\tilde{R}$. The operations used in this algorithm are additions, multiplications, subtractions, divisions, square roots, and absolute values. So if we apply this algorithm to a matrix with entries in $\mathcal{S}_{\mathrm{e}}$, the entries of all the matrices generated during the iteration process also belong to $\mathcal{S}_{\mathrm{e}}$ by Proposition 5.3.

In theory we should run the row-cyclic Kogbetliantz algorithm forever in order to produce a path of exact SVDs of $\tilde{A}$. However, since we are only interested in the asymptotic behavior of the singular values and the entries of the singular vectors of $\tilde{A}$, we may stop the iteration process after a finite number of sweeps.

Let $\tilde{S}_k$, $\tilde{U}_k$, and $\tilde{V}_k$ be the matrix-valued functions that are computed in the kth step of the algorithm. Let $\tilde{\Delta}_p$ be the diagonal matrix-valued function obtained by removing the off-diagonal entries of $\tilde{S}_{pN}$ (where $N = \frac{n(n-1)}{2}$ is the number of iterations per sweep), making all diagonal entries nonnegative and arranging them in descending order, and adding $m - n$ zero rows (cf. the transformations used to go from S to S_A in the explanation of Kogbetliantz's algorithm given above). Let $\tilde{X}_p$ and $\tilde{Y}_p$ be the matrix-valued functions obtained by applying the corresponding transformations to $\tilde{U}_{pN}$ and $\tilde{V}_{pN}$, respectively. If we define a matrix-valued function $\tilde{C}_p = \tilde{X}_p \tilde{\Delta}_p \tilde{Y}_p^T$, we have a path of *exact* SVDs of $\tilde{C}_p$ on some interval $[L, \infty)$. This means that we may stop the iteration process as soon as

$$\mathcal{F}(\tilde{A}, N, s) \sim \tilde{C}_p(s),\ s \to \infty \tag{33}$$

for some $N \in \mathbb{R}_0^{m\times n}$. Note that eventually this condition will always be satisfied due to the fact that Kogbetliantz's SVD algorithm is globally convergent and—for triangular matrices—also quadratically convergent if p is large enough, and due to the fact that the entries of $\tilde{A}$—to which the entries of $\tilde{C}_p$ should converge—are not identically zero since they have a finite dominant exponent.

Let $\tilde{U}\tilde{S}\tilde{V}^T$ be a path of approximate SVDs of $\tilde{A}$ on some interval $[L, \infty)$ that was obtained by the procedure given above. Since we have performed a *finite* number of

elementary operations on the entries of $\tilde{A}$, the entries of $\tilde{U}$, $\tilde{S}$, and $\tilde{V}$ belong to $\mathcal{S}_{\mathrm{e}}$. We have

$$\mathcal{F}(\tilde{A}, N, s) \;\sim\; \tilde{U}(s)\,\tilde{\Sigma}(s)\,\tilde{V}^T(s)\;,\; s \to \infty \tag{34}$$

for some $N \in \mathbb{R}_0^{m\times n}$. Furthermore,

$$\tilde{U}^T(s)\,\tilde{U}(s) = I_m \qquad \text{for all } s \geq L, \tag{35}$$

$$\tilde{V}^T(s)\,\tilde{V}(s) = I_n \qquad \text{for all } s \geq L\,. \tag{36}$$

The diagonal entries of $\tilde{\Sigma}$ and the entries of $\tilde{U}$ and $\tilde{V}$ belong to $\mathcal{S}_{\mathrm{e}}$ and are thus asymptotically equivalent to an exponential in the neighborhood of ∞ by Lemma 5.2. Define $\tilde{\sigma}_i = \tilde{\Sigma}_{ii}$ for $i = 1, 2, \ldots, r$.

Now we apply the reverse mapping $\mathcal{R}$ in order to obtain a max-algebraic SVD of A. If we define

$$\Sigma = \mathcal{R}(\tilde{\Sigma})\,,\; U = \mathcal{R}(\tilde{U})\,,\; V = \mathcal{R}(\tilde{V}),\; \text{and}\;\; \sigma_i = (\Sigma)_{ii} = \mathcal{R}(\tilde{\sigma}_i) \text{ for all } i\,,$$

then Σ is a max-algebraic diagonal matrix and U and V have signed entries. If we apply the reverse mapping $\mathcal{R}$ to (34)–(36), we get

$$A \,\nabla\, U \otimes \Sigma \otimes V^T,\;\; U^T \otimes U \,\nabla\, E_m,\; \text{and}\; V^T \otimes V \,\nabla\, E_n.$$

The $\tilde{\sigma}_i$'s are nonnegative in $[L, \infty)$ and therefore we have $\sigma_i \in \mathbb{R}_\varepsilon$ for all i. Since the $\tilde{\sigma}_i$'s are ordered in $[L, \infty)$, their dominant exponents are also ordered. Hence, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$.

We have $\|\tilde{A}(s)\|_{\mathrm{F}} \sim \gamma e^{cs}$, $s \to \infty$ for some $\gamma > 0$ since $c = \|A\|_{\oplus}$ is the largest exponent that appears in the entries of $\tilde{A}$. Hence, $\mathcal{R}(\|\tilde{A}\|_{\mathrm{F}}) = c = \|A\|_{\oplus}$.

If $P \in \mathbb{R}^{m\times n}$, then $\dfrac{1}{\sqrt{n}}\,\|P\|_{\mathrm{F}} \leq \|P\|_2 \leq \|P\|_{\mathrm{F}}$. As a consequence we have

$$\frac{1}{\sqrt{n}}\,\|\tilde{A}\|_{\mathrm{F}} \leq \|\tilde{A}\|_2 \leq \|\tilde{A}\|_{\mathrm{F}} \qquad \text{for all } s \geq L\,.$$

Since $\tilde{\sigma}_1(s) \sim \|\tilde{A}(s)\|_2$, $s \to \infty$ and since the mapping $\mathcal{R}$ preserves the order, this leads to $\|A\|_{\oplus} \leq \sigma_1 \leq \|A\|_{\oplus}$ and, consequently, $\sigma_1 = \|A\|_{\oplus}$.

Case 2. Not all the a_{ij}'s have a finite max-absolute value.

First we construct a sequence $\{A_l\}_{l=1}^{\infty}$ of $m \times n$ matrices such that

$$(A_l)_{ij} = \begin{cases} a_{ij} & \text{if } |a_{ij}|_{\oplus} \neq \varepsilon\,, \\ \|A\|_{\oplus} - l & \text{if } |a_{ij}|_{\oplus} = \varepsilon\,, \end{cases}$$

for all i, j. So the entries of the matrices A_l are finite and $\|A\|_{\oplus} = \|A_l\|_{\oplus}$ for all $l \in \mathbb{N}_0$. Furthermore, $\lim_{l\to\infty} A_l = A$.

Now we construct the sequence $\{\tilde{A}_l\}_{l=1}^{\infty}$ with $\tilde{A}_l = \mathcal{F}(A_l, M, \cdot)$ for $l = 1, 2, 3, \ldots$ with $M \in \mathbb{R}^{m\times n}$ and $m_{ij} = 1$ for all i, j. We compute a path of approximate SVDs $\tilde{U}_l\,\tilde{\Sigma}_l\,\tilde{V}_l^T$ of each $\tilde{A}_l$ using the method of Case 1 of this proof.

In general, it is possible that for some of the entries of the $\tilde{U}_l$'s and the $\tilde{V}_l$'s the sequence of the dominant exponents and the sequence of the corresponding coefficients have more than one accumulation point (since if two or more singular values coincide, the corresponding left and right singular vectors are not uniquely defined). However,

since we use a fixed computation scheme (the row-cyclic Kogbetliantz algorithm), all the sequences will have exactly one accumulation point. So some of the dominant exponents will reach a finite limit as l goes to ∞, while the other dominant exponents will tend to $-\infty$. If we take the reverse mapping $\mathcal{R}$, we get a sequence of max-algebraic SVDs $\{U_l \otimes \Sigma_l \otimes V_l^T\}_{l=1}^{\infty}$, where some of the entries, viz., those that correspond to dominant exponents that tend to $-\infty$, tend to ε as l goes to ∞.

If we define

$$U = \lim_{l\to\infty} U_l, \ \Sigma = \lim_{l\to\infty} \Sigma_l, \ \text{and} \ V = \lim_{l\to\infty} V_l,$$

then we have

$$A \nabla U \otimes \Sigma \otimes V^T, \ \ U^T \otimes U \nabla E_m, \ \text{and} \ V^T \otimes V \nabla E_n.$$

Since the diagonal entries of all the Σ_l's belong to $\mathbb{R}_\varepsilon$ and are ordered, the diagonal entries of Σ also belong to $\mathbb{R}_\varepsilon$ and are also ordered. Furthermore, $(\Sigma)_{11} = \|A\|_\oplus$ since $(\Sigma_l)_{11} = \|A\|_\oplus$ for all l. Hence, $U \otimes \Sigma \otimes V^T$ is a max-algebraic SVD of A. □

Remark 5.8. Now we explain why we have distinguished between two different cases in the proof of Theorem 5.7.

If there exist indices i and j such that $a_{ij} = \varepsilon$, then $\tilde{a}_{ij}(s) = 0$ for all $s \in \mathbb{R}_0^+$, which means that we cannot guarantee that condition (33) will be satisfied after a finite number of sweeps. This is why we make a distinction between the case where all the entries of A are finite and the case where at least one entry of A is equal to ε.

Let us now show that we do not have to take special precautions if $\tilde{A}$ has singular values that are identically zero in the neighborhood of ∞. If $\tilde{\Psi}$ is a real matrix-valued function that is analytic in some interval $J \subseteq \mathbb{R}$, then the rank of $\tilde{\Psi}$ is constant in J except in some isolated points where the rank drops [13]. If the rank of $\tilde{\Psi}(s)$ is equal to ρ for all $s \in J$ except for some isolated points, then we say that the *generic rank* of $\tilde{\Psi}$ in J is equal to ρ. The entries of all the matrix-valued functions created in the row-cyclic Kogbetliantz algorithm when applied to $\tilde{A}$ are real and analytic in some interval $[L^*, \infty)$. Furthermore, for a fixed value of s the matrices $\tilde{A}(s)$, $\tilde{R}(s)$, $\tilde{S}_1(s)$, $\tilde{S}_2(s)$, ... all have the same rank since they are related by orthogonal transformations. So if ρ is the generic rank of $\tilde{A}$ in $[L^*, \infty)$, then the generic rank of $\tilde{R}$, $\tilde{S}_1$, $\tilde{S}_2$, ... in $[L^*, \infty)$ is also equal to ρ. If $\rho < n$, then the $n-\rho$ smallest singular values of $\tilde{R}$ will be identically zero in $[L^*, \infty)$. However, since $\tilde{R}$, $\tilde{S}_N$, $\tilde{S}_{2N}$, ... are triangular matrices, they have at least $n - \rho$ diagonal entries that are identically zero in $[L^*, \infty)$ since otherwise their generic rank would be greater than ρ. In fact this also holds for $\tilde{S}_1$, $\tilde{S}_2$, ... since these matrix-valued functions are hierarchically triangular, i.e., block triangular such that the diagonal blocks are again block triangular, etc. [17]. Furthermore, if k is large enough, diagonal entries do not change their affiliation any more, i.e., if a diagonal entry corresponds to a specific singular value in the kth iteration, then it will also correspond to that singular value in all the next iterations. Since the diagonal entries of $\tilde{S}_k$ are asymptotically equivalent to an exponential in the neighborhood of ∞, this means that at least $n-\rho$ diagonal entries (with a fixed position) of $\tilde{S}_k$, $\tilde{S}_{k+1}$, ... will be identically zero in some interval $[L, \infty) \subseteq [L^*, \infty)$ if k is large enough. Hence, we do not have to take special precautions if $\tilde{A}$ has singular values that are identically zero in the neighborhood of ∞ since convergence to these singular values in a finite number of iteration steps is guaranteed.

For inner products of two different columns of $\tilde{U}$ there are no problems either: these inner products are equal to zero by construction since the matrix-valued function

$\tilde{U}_k$ is orthogonal on $[L, \infty)$ for all $k \in \mathbb{N}$. This also holds for inner products of two different columns of $\tilde{V}$. □

If $U\Sigma V^T$ is an SVD of a matrix $A \in \mathbb{R}^{m\times n}$, then we have $\sigma_1 = (\Sigma)_{11} = \|A\|_2$. However, in $\mathbb{S}_{\max}$ the balances $A \nabla U \otimes \Sigma \otimes V^T$, $U^T \otimes U \nabla E_m$, and $V^T \otimes V \nabla E_n$, where Σ is a diagonal matrix with entries in $\mathbb{R}_\varepsilon$ and where the entries of U and V are signed, do not always imply that $(\Sigma)_{11} = \|A\|_\oplus$ [10]. Therefore, we have included the extra condition $\sigma_1 = \|A\|_\oplus$ in the definition of the max-algebraic SVD.

Using a reasoning that is similar to the one that has been used at the end of section 5.2, we can show that only permuted max-algebraic diagonal matrices with entries in $\mathbb{R}_\varepsilon$ have a max-algebraic SVD with entries in $\mathbb{R}_\varepsilon$ [7, 10].

For properties of the max-algebraic SVD and for a possible application of this decomposition in a method to solve the identification problem for max-linear DESs, the interested reader is referred to [7, 10]. In [7] we have also proposed some possible extensions of the definitions of the max-algebraic QR decomposition and the max-algebraic SVD.

The proof technique that has been used in this section essentially consists of applying an algorithm from linear algebra to a matrix with entries in $\mathcal{S}_e$. This proof technique can also be used to prove the existence of many other max-algebraic matrix decompositions: it can easily be adapted to prove the existence of a max-algebraic eigenvalue decomposition for symmetric matrices (by using the Jacobi algorithm for the computation of the eigenvalue decomposition of a real symmetric matrix), a max-algebraic LU decomposition, a max-algebraic Schur decomposition, a max-algebraic Hessenberg decomposition, and so on.

6. A worked example of the max-algebraic QR decomposition and the max-algebraic SVD. Now we give an example of the computation of a max-algebraic QR decomposition and a max-algebraic SVD of a matrix using the mapping $\mathcal{F}$.

Example 6.1. Consider the matrix

$$A = \begin{bmatrix} \ominus 0 & 3^\bullet & \ominus(-1) \\ 1 & \ominus(-2) & \varepsilon \end{bmatrix}.$$

Let us first compute a max-algebraic QR decomposition of A using the mapping $\mathcal{F}$. Let $M = \left[\begin{smallmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{smallmatrix}\right]$ and define $\tilde{A} = \mathcal{F}(A, M, \cdot)$. Hence,

$$\tilde{A}(s) = \begin{bmatrix} -1 & e^{3s} & -e^{-s} \\ e^s & -e^{-2s} & 0 \end{bmatrix} \qquad \text{for all } s \in \mathbb{R}_0^+.$$

If we use the Givens QR algorithm, we get a path of QR decompositions $\tilde{Q}\tilde{R}$ of $\tilde{A}$ with

$$\tilde{Q}(s) = \begin{bmatrix} \dfrac{-e^{-s}}{\sqrt{1+e^{-2s}}} & \dfrac{-1}{\sqrt{1+e^{-2s}}} \\ \dfrac{1}{\sqrt{1+e^{-2s}}} & \dfrac{-e^{-s}}{\sqrt{1+e^{-2s}}} \end{bmatrix},$$

$$\tilde{R}(s) = \begin{bmatrix} e^s\sqrt{1+e^{-2s}} & \dfrac{-e^{2s}-e^{-2s}}{\sqrt{1+e^{-2s}}} & \dfrac{e^{-2s}}{\sqrt{1+e^{-2s}}} \\ 0 & \dfrac{-e^{3s}+e^{-3s}}{\sqrt{1+e^{-2s}}} & \dfrac{e^{-s}}{\sqrt{1+e^{-2s}}} \end{bmatrix}$$

for all $s \in \mathbb{R}_0^+$. Hence,

$$\tilde{Q}(s) \sim \begin{bmatrix} -e^{-s} & -1 \\ 1 & -e^{-s} \end{bmatrix}, \ s \to \infty,$$

$$\tilde{R}(s) \sim \begin{bmatrix} e^{s} & -e^{2s} & e^{-2s} \\ 0 & -e^{3s} & e^{-s} \end{bmatrix}, \ s \to \infty.$$

If we define $Q = \mathcal{R}(\tilde{Q})$ and $R = \mathcal{R}(\tilde{R})$, we obtain

$$Q = \begin{bmatrix} \ominus(-1) & \ominus 0 \\ 0 & \ominus(-1) \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 1 & \ominus 2 & -2 \\ \varepsilon & \ominus 3 & -1 \end{bmatrix}.$$

We have

$$Q \otimes R = \begin{bmatrix} \ominus 0 & 3 & \ominus(-1) \\ 1 & 2^{\bullet} & (-2)^{\bullet} \end{bmatrix} \nabla A,$$

$$Q^T \otimes Q = \begin{bmatrix} 0 & (-1)^{\bullet} \\ (-1)^{\bullet} & 0 \end{bmatrix} \nabla E_2,$$

and $\|R\|_{\oplus} = 3 = \|A\|_{\oplus}$.

Let us now compute a max-algebraic SVD of A. Since $\tilde{A}$ is a 2×3 matrix-valued function, we can compute a path of SVDs $\tilde{U}\tilde{\Sigma}\tilde{V}^T$ of $\tilde{A}$ analytically, e.g., via the eigenvalue decomposition of $\tilde{A}^T\tilde{A}$ (see [14, 32]). This yields[1]

$$\tilde{U}(s) \sim \begin{bmatrix} 1 & 2\,e^{-5s} \\ -2\,e^{-5s} & 1 \end{bmatrix}, \ s \to \infty,$$

$$\tilde{\Sigma}(s) \sim \begin{bmatrix} e^{3s} & 0 & 0 \\ 0 & e^{s} & 0 \end{bmatrix}, \ s \to \infty,$$

$$\tilde{V}(s) \sim \begin{bmatrix} -e^{-3s} & 1 & e^{-7s} \\ 1 & e^{-3s} & e^{-4s} \\ -e^{-4s} & -2\,e^{-7s} & 1 \end{bmatrix}, \ s \to \infty.$$

If we apply the reverse mapping $\mathcal{R}$, we get a max-algebraic SVD $U \otimes \Sigma \otimes V^T$ of A with

$$U = \mathcal{R}(\tilde{U}) = \begin{bmatrix} 0 & -5 \\ \ominus(-5) & 0 \end{bmatrix},$$

$$\Sigma = \mathcal{R}(\tilde{\Sigma}) = \begin{bmatrix} 3 & \varepsilon & \varepsilon \\ \varepsilon & 1 & \varepsilon \end{bmatrix},$$

$$V = \mathcal{R}(\tilde{V}) = \begin{bmatrix} \ominus(-3) & 0 & -7 \\ 0 & -3 & -4 \\ \ominus(-4) & \ominus(-7) & 0 \end{bmatrix}.$$

We have

$$U \otimes \Sigma \otimes V^T = \begin{bmatrix} \ominus 0 & 3 & \ominus(-1) \\ 1 & (-2)^{\bullet} & (-6)^{\bullet} \end{bmatrix} \nabla A,$$

[1] We have used the symbolic computation tool MAPLE to compute a path of SVDs $\tilde{U}\tilde{\Sigma}\tilde{V}^T$ of $\tilde{A}$. However, since the full expressions for the entries of $\tilde{U}$, $\tilde{S}$, and $\tilde{V}$ are too long and too intricate to display here, we only give the dominant exponentials.

$$U^T \otimes U = \begin{bmatrix} 0 & (-5)^\bullet \\ (-5)^\bullet & 0 \end{bmatrix} \nabla E_2,$$

$$V^T \otimes V = \begin{bmatrix} 0 & (-3)^\bullet & (-4)^\bullet \\ (-3)^\bullet & 0 & (-7)^\bullet \\ (-4)^\bullet & (-7)^\bullet & 0 \end{bmatrix} \nabla E_3,$$

and $\sigma_1 = 3 = \|A\|_\oplus \geq 1 = \sigma_2$. □

Another example of the computation of a max-algebraic SVD can be found in [7, 10].

Remark 6.2. In [7] we have shown that the max-algebraic QR decomposition and the max-algebraic SVD of a matrix can also be computed by solving an extended linear complementarity problem (ELCP)—which is a kind of mathematical programming problem. Although it would lead us to far to explain this procedure in detail, we shall now give a brief outline of how the equations that appear in the definition of the max-algebraic QR decomposition and the max-algebraic SVD can be transformed into a system of multivariate max-algebraic polynomial equalities.

Consider the equation $A \nabla Q \otimes R$. If we extract the max-positive and the max-negative parts of each matrix, we obtain

$$A^\oplus \ominus A^\ominus \nabla (Q^\oplus \ominus Q^\ominus) \otimes (R^\oplus \ominus R^\ominus)$$

or

$$A^\oplus \ominus A^\ominus \nabla Q^\oplus \otimes R^\oplus \ominus Q^\oplus \otimes R^\ominus \ominus Q^\ominus \otimes R^\oplus \oplus Q^\ominus \otimes R^\ominus.$$

By Proposition 3.5 this can be rewritten as

$$A^\oplus \oplus Q^\oplus \otimes R^\ominus \oplus Q^\ominus \otimes R^\oplus \nabla A^\ominus \oplus Q^\oplus \otimes R^\oplus \oplus Q^\ominus \otimes R^\ominus.$$

Both sides of this balance are signed. So by Proposition 3.6 we may replace the balance by an equality. If we introduce a matrix T of auxiliary variables, we obtain

$$A^\oplus \oplus Q^\oplus \otimes R^\ominus \oplus Q^\ominus \otimes R^\oplus = T, \tag{37}$$

$$A^\ominus \oplus Q^\oplus \otimes R^\oplus \oplus Q^\ominus \otimes R^\ominus = T. \tag{38}$$

If we write out the max-algebraic matrix multiplications in (37) and if we transfer the entries of T to the opposite side, we get

$$\begin{aligned} & a_{ij}^\oplus \otimes {t_{ij}}^{\otimes^{-1}} \oplus \bigoplus_{k=1}^{m} q_{ik}^\oplus \otimes r_{kj}^\ominus \otimes {t_{ij}}^{\otimes^{-1}} \\ & \oplus \bigoplus_{k=1}^{m} q_{ik}^\ominus \otimes r_{kj}^\oplus \otimes {t_{ij}}^{\otimes^{-1}} = 0 \qquad \text{for all } i, j. \end{aligned} \tag{39}$$

Equation (38) can be rewritten in a similar way. The condition $Q^T \otimes Q \nabla E_m$ also leads to similar equations.

The condition that the entries of Q and R should be signed can be written as

$$q_{ij}^\oplus \otimes q_{ij}^\ominus = \varepsilon \qquad \text{for all } i, j, \tag{40}$$

$$r_{ij}^\oplus \otimes r_{ij}^\ominus = \varepsilon \qquad \text{for all } i, j. \tag{41}$$

The condition $\|R\|_{\oplus} = \|A\|_{\oplus}$ is equivalent to

$$\bigoplus_{i=1}^{m} \bigoplus_{j=1}^{n} \left(r_{ij}^{\oplus} \oplus r_{ij}^{\ominus} \right) = \|A\|_{\oplus} \qquad \text{for all } i, j. \tag{42}$$

So if we combine all equations of the form (39)–(42), we obtain a system of multivariate max-algebraic polynomial equalities of the following form:

> Given l integers $m_1,\ m_2,\ \ldots,\ m_l \in \mathbb{N}_0$ and real numbers $a_{ki},\ b_k,$ and c_{kij} for $k = 1, 2, \ldots, l,\ i = 1, 2, \ldots, m_l,$ and $j = 1, 2, \ldots, r,$ find $x \in \mathbb{R}_{\varepsilon}^{r}$ such that
>
> $$\bigoplus_{i=1}^{m_l} a_{ki} \otimes \bigotimes_{j=1}^{r} x_j^{\otimes^{c_{kij}}} = b_k \quad \text{for } k = 1, 2, \ldots, l\,,$$
>
> or show that no such x exists.

Here the vector x contains the max-positive and max-negative parts of the entries of Q and R and the auxiliary variables.

Using a similar reasoning we can also show that the equations that appear in the definition of the max-algebraic SVD also lead to a system of multivariate max-algebraic polynomial equalities.

In [7, 9] we have shown that a system of multivariate max-algebraic polynomial equalities can be rewritten as a mathematical programming problem of the following form:

> Given two matrices $A \in \mathbb{R}^{p \times r}$, $B \in \mathbb{R}^{q \times r}$, two vectors $c \in \mathbb{R}^p$, $d \in \mathbb{R}^q$, and s subsets $\phi_1,\ \phi_2,\ \ldots,\ \phi_s$ of $\{1, 2, \ldots, p\}$, find $x \in \mathbb{R}^r$ such that
>
> $$\sum_{j=1}^{s} \prod_{i \in \phi_j} (Ax - c)_i = 0$$
>
> subject to $Ax \geq c$ and $Bx = d$, or show that no such x exists.

This problem is called an *extended linear complementarity problem* (ELCP). In [7, 8] we have developed an algorithm to find all solutions of a general ELCP. However, the execution time of this algorithm increases exponentially as the number of equations and variables of the ELCP increases. Furthermore, in [7, 8] we have shown that the general ELCP is an NP-hard problem. As a consequence, the ELCP approach can only be used to compute max-algebraic QR decompositions and max-algebraic SVDs of small-sized matrices. So there certainly is a need for efficient algorithms to compute max-algebraic QR decompositions and SVDs: this will be one of the most important topics for further research. An important question is whether we can develop efficient algorithms for special classes of matrices, e.g., is it possible to come up with more efficient algorithms by making use of the nonzero structure (sparsity, bandedness, ...) of the matrix? □

7. Conclusions and future research. In this paper we have tried to fill one of the gaps in the theory of the (symmetrized) max-plus algebra by showing that there exist max-algebraic analogues of many fundamental matrix decompositions from linear algebra.

We have established a link between a ring of real functions (with addition and multiplication as basic operations) and the symmetrized max-plus algebra. Next we have introduced a class of functions that are analytic and that can be written as a

sum or a series of exponentials in a neighborhood of ∞. This class is closed under basic operations such as additions, subtractions, multiplications, divisions, powers, square roots, and absolute values. This fact has then been used to prove the existence of a QR decomposition and an SVD of a matrix in the symmetrized max-plus algebra. These decompositions are max-algebraic analogues of basic matrix decompositions from linear algebra. The proof technique that has been used to prove the existence of these max-algebraic matrix decompositions can also be used to prove the existence of max-algebraic analogues of other real matrix decompositions from linear algebra such as the LU decomposition, the Hessenberg decomposition, the eigenvalue decomposition (for symmetric matrices), the Schur decomposition, and so on.

In [7, 10] we have introduced a further extension of the symmetrized max-plus algebra: the max-complex structure $\mathbb{T}_{\max}$, which corresponds to a ring of complex functions (with addition and multiplication as basic operations). We could also define max-algebraic matrix decompositions in $\mathbb{T}_{\max}$. These decompositions would then be analogues of matrix decompositions from linear algebra for complex matrices (such as the eigenvalue decomposition or the Jordan decomposition).

Topics for future research are as follows: further investigation of the properties of the max-algebraic matrix decompositions that have been introduced in this paper, development of efficient algorithms to compute these max-algebraic matrix decompositions, investigation of the computational complexity of computing max-algebraic matrix decompositions (in general and for special classes of matrices), and application of the max-algebraic SVD and other max-algebraic matrix decompositions in the system theory for max-linear discrete event systems.

Appendix A. Proof of Lemma 5.2. In this section we show that functions that belong to the class $\mathcal{S}_{\mathrm{e}}$ are asymptotically equivalent to an exponential in the neighborhood of ∞. We shall use the following lemma.

LEMMA A.1. *If $f \in \mathcal{S}_{\mathrm{e}}$ is a series with a nonpositive dominant exponent, i.e., if there exists a positive real number K such that $f(x) = \sum_{i=0}^{\infty} \alpha_i e^{a_i x}$ for all $x \geq K$ with $\alpha_i \in \mathbb{R}$, $a_i \in \mathbb{R}^-$, $a_i > a_{i+1}$ for all i, $\lim_{i\to\infty} a_i = \varepsilon$, and where the series converges absolutely for every $x \geq K$, then the series $\sum_{i=0}^{\infty} \alpha_i e^{a_i x}$ converges uniformly in $[K, \infty)$.*

Proof. If $x \geq K$, then we have $e^{a_i x} \leq e^{a_i K}$ for all $i \in \mathbb{N}_0$ since $a_i \leq 0$ for all i. Hence, $|\alpha_i e^{a_i x}| \leq |\alpha_i e^{a_i K}|$for all $x \geq K$ and for all $i \in \mathbb{N}_0$. We already know that $\sum_{i=1}^{\infty} |\alpha_i e^{a_i K}|$ converges. Now we can apply the Weierstrass M-test (see [19, 24]). As a consequence, the series $\sum_{i=1}^{\infty} \alpha_i e^{a_i x}$converges uniformly in $[K, \infty)$. □

Proof (proof of Lemma 5.2). If $f \in \mathcal{S}_{\mathrm{e}}$, then there exists a positive real number K such that $f(x) = \sum_{i=0}^{n} \alpha_i e^{a_i x}$ for all $x \geq K$ with $n \in \mathbb{N} \cup \{\infty\}$, $\alpha_i \in \mathbb{R}_0$, and $a_i \in \mathbb{R}_\varepsilon$ for all i. If $n = \infty$, then f is a series that converges absolutely in $[K, \infty)$.

If $a_0 = \varepsilon$, then there exists a real number K such that $f(x) = 0$ for all $x \geq K$ and then we have $f(x) \sim 0 = 1 \cdot e^{\varepsilon x}$, $x \to \infty$ by Definition 2.2.

If $n = 1$, then $f(x) = \alpha_0 e^{a_0 x}$ and thus $f(x) \sim \alpha_0 e^{a_0 x}$, $x \to \infty$ with $\alpha_0 \in \mathbb{R}_0$ and $a_0 \in \mathbb{R}_\varepsilon$.

From now on we assume that $n > 1$ and $a_0 \neq \varepsilon$. Then we can rewrite $f(x)$ as

$$f(x) = \alpha_0 e^{a_0 x} \left(1 + \sum_{i=1}^{n} \frac{\alpha_i}{\alpha_0} \, e^{(a_i - a_0)x} \right) = \alpha_0 e^{a_0 x} (\, 1 + p(x) \,)$$

with $p(x) = \sum_{i=1}^{n} \gamma_i e^{c_i x}$, where $\gamma_i = \dfrac{\alpha_i}{\alpha_0} \in \mathbb{R}_0$ and $c_i = a_i - a_0 < 0$ for all i. Note

that $p \in \mathcal{S}_\mathrm{e}$ and p has a negative dominant exponent. Since $c_i < 0$ for all i, we have

$$\lim_{x\to\infty} p(x) = \lim_{x\to\infty} \left(\sum_{i=1}^{n} \gamma_i e^{c_i x} \right) = \sum_{i=1}^{n} \left(\lim_{x\to\infty} \gamma_i e^{c_i x} \right) = 0. \tag{43}$$

If $n = \infty$, then the series $\sum_{i=1}^{\infty} \gamma_i e^{c_i x}$ converges uniformly in $[K, \infty)$ by Lemma A.1. As a consequence, we may also interchange the summation and the limit in (43) if $n = \infty$ (cf. [19]).

Now we have

$$\lim_{x\to\infty} \frac{f(x)}{\alpha_0 e^{a_0 x}} = \lim_{x\to\infty} \frac{\alpha_0 e^{a_0 x}(1 + p(x))}{\alpha_0 e^{a_0 x}} = \lim_{x\to\infty} (1 + p(x)) = 1,$$

and thus $f(x) \sim \alpha_0 e^{a_0 x}$, $x \to \infty$ where $\alpha_0 \in \mathbb{R}_0$ and $a_0 \in \mathbb{R}$. ∎

Appendix B. Proof of Proposition 5.3. In this section we show that $\mathcal{S}_\mathrm{e}$ is closed under elementary operations such as additions, multiplications, subtractions, divisions, square roots, and absolute values.

Proof (proof of Proposition 5.3). If f and g belong to $\mathcal{S}_\mathrm{e}$, then we may assume without loss of generality that the domains of definition of f and g coincide, since we can always restrict the functions f and g to $\operatorname{dom} f \cap \operatorname{dom} g$ and since the restricted functions also belong to $\mathcal{S}_\mathrm{e}$.

Since f and g belong to $\mathcal{S}_\mathrm{e}$, there exists a positive real number K such that

$$f(x) = \sum_{i=0}^{n} \alpha_i e^{a_i x} \quad \text{and} \quad g(x) = \sum_{j=0}^{m} \beta_j e^{b_j x} \qquad \text{for all } x \geq K$$

with $m, n \in \mathbb{N} \cup \{\infty\}$, $\alpha_i, \beta_j \in \mathbb{R}_0$, and $a_i, b_j \in \mathbb{R}_\varepsilon$ for all i, j. If m or n is equal to ∞, then the corresponding series converges absolutely in $[K, \infty)$.

We may assume without loss of generality that both m and n are equal to ∞. If m or n are finite, then we can always add dummy terms of the form $0 \cdot e^{\varepsilon x}$ to $f(x)$ or $g(x)$. Afterwards we can remove terms of the form $re^{\varepsilon x}$ with $r \in \mathbb{R}$ to obtain an expression with nonzero coefficients and decreasing exponents. So from now on we assume that both f and g are absolute convergent series of exponentials.

If $a_0 = \varepsilon$, then we have $f(x) = 0$ for all $x \geq K$, which means that $|f(x)| = 0$ for all $x \geq K$. So if $a_0 = \varepsilon$, then $|f|$ belongs to $\mathcal{S}_\mathrm{e}$.

If $a_0 \neq \varepsilon$, then there exists a real number $L \geq K$ such that either $f(x) > 0$ or $f(x) < 0$ for all $x \geq L$ since $f(x) \sim \alpha_0 e^{a_0 x}$, $x \to \infty$ with $\alpha_0 \neq 0$ by Lemma 5.2. Hence, either $|f(x)| = f(x)$ or $|f(x)| = -f(x)$ for all $x \geq L$. So in this case $|f|$ also belongs to $\mathcal{S}_\mathrm{e}$.

Since f and g are analytic in $[K, \infty)$, the functions ρf, $f + g$, $f - g$, $f \cdot g$, and f^l are also analytic in $[K, \infty)$ for any $\rho \in \mathbb{R}$ and any $l \in \mathbb{N}$.

Now we prove that these functions can be written as a sum of exponentials or as an absolutely convergent series of exponentials.

Consider an arbitrary $\rho \in \mathbb{R}$. If $\rho = 0$, then $\rho f(x) = 0$ for all $x \geq K$ and thus $\rho f \in \mathcal{S}_\mathrm{e}$.

If $\rho \neq 0$, then we have $\rho f(x) = \sum_{i=0}^{\infty} (\rho \alpha_i) e^{a_i x}$. The series $\sum_{i=0}^{\infty} (\rho \alpha_i) e^{a_i x}$ also converges absolutely in $[K, \infty)$ and has the same exponents as $f(x)$. Hence, $\rho f \in \mathcal{S}_\mathrm{e}$.

The sum function $f + g$ is a series of exponentials since

$$f(x) + g(x) = \sum_{i=0}^{\infty} \alpha_i e^{a_i x} + \sum_{j=0}^{\infty} \beta_j e^{b_j x}.$$

Furthermore, this series converges absolutely for every $x \geq K$. Therefore, the sum of the series does not change if we rearrange the terms [19]. So $f(x) + g(x)$ can be written in the form of Definition 5.1 by reordering the terms, adding up terms with equal exponents and removing terms of the form $re^{\varepsilon x}$ with $r \in \mathbb{R}$, if there are any. If the result is a series, then the sequence of exponents is decreasing and unbounded from below. So $f + g \in \mathcal{S}_{\mathrm{e}}$.

Since $f - g = f + (-1)g$, the function $f - g$ also belongs to $\mathcal{S}_{\mathrm{e}}$.

The series corresponding to f and g converge absolutely for every $x \geq K$. Therefore, their Cauchy product will also converge absolutely for every $x \geq K$ and it will be equal to fg [19]:

$$f(x)g(x) = \sum_{i=0}^{\infty} \sum_{j=0}^{i} \alpha_j \beta_{i-j} e^{(a_j + b_{i-j})x} \qquad \text{for all } x \geq K\,.$$

Using the same procedure as for $f + g$, we can also write this product in the form (24) or (25). So $fg \in \mathcal{S}_{\mathrm{e}}$.

Let $l \in \mathbb{N}$. If $l = 0$, then $f^l = 0 \in \mathcal{S}_{\mathrm{e}}$ and if $l = 1$, then $f^l = f \in \mathcal{S}_{\mathrm{e}}$. If $l > 1$, we can make repeated use of the fact that $fg \in \mathcal{S}_{\mathrm{e}}$ if $f, g \in \mathcal{S}_{\mathrm{e}}$ to prove that f^l also belongs to $\mathcal{S}_{\mathrm{e}}$.

If there exists a real number P such that $f(x) \neq 0$ for all $x \geq P$, then $\frac{1}{f}$ and $\frac{g}{f}$ are defined and analytic in $[P, \infty)$. Note that we may assume without loss of generality that $P \geq K$. Furthermore, since the function f restricted to the interval $[P, \infty)$ also belongs to $\mathcal{S}_{\mathrm{e}}$, we may assume without loss of generality that the domain of definition of f is $[P, \infty)$.

If $f(x) \neq 0$ for all $x \geq P$, then we have $a_0 \neq \varepsilon$. As a consequence, we can rewrite $f(x)$ as

$$f(x) = \sum_{i=0}^{\infty} \alpha_i e^{a_i x} = \alpha_0 e^{a_0 x} \left(1 + \sum_{i=1}^{\infty} \frac{\alpha_i}{\alpha_0}\ e^{(a_i - a_0)x} \right) = \alpha_0 e^{a_0 x} (\, 1 + p(x)\,)$$

with $p(x) = \sum_{i=1}^{\infty} \gamma_i e^{c_i x}$, where $\gamma_i = \frac{\alpha_i}{\alpha_0} \in \mathbb{R}_0$ and $c_i = a_i - a_0 < 0$ for all i. Note that p is defined in $[P, \infty)$, that $p \in \mathcal{S}_{\mathrm{e}}$, and that p has a negative dominant exponent.

If $c_1 = \varepsilon$, then $p(x) = 0$ and $\frac{1}{f(x)} = \frac{1}{\alpha_0} e^{-a_0 x}$ for all $x \geq P$. Hence, $\frac{1}{f} \in \mathcal{S}_{\mathrm{e}}$.

Now assume that $c_1 \neq \varepsilon$. Since $\{c_i\}_{i=1}^{\infty}$ is a nonincreasing sequence of negative numbers with $\lim_{i \to \infty} c_i = \varepsilon = -\infty$ and since p converges uniformly in $[P, \infty)$ by Lemma A.1, we have $\lim_{x \to \infty} p(x) = 0$ (cf. (43)). So $|\, p(x)\,|$ will be less than one if x is large enough, say if $x \geq M$. If we use the Taylor series expansion of $\frac{1}{1+x}$, we obtain

$$\frac{1}{1 + p(x)} = \sum_{k=0}^{\infty} (-1)^k p^k(x) \qquad \text{if } |p(x)| < 1\,. \tag{44}$$

We already know that $p \in \mathcal{S}_{\mathrm{e}}$. Hence, $p^k \in \mathcal{S}_{\mathrm{e}}$ for all $k \in \mathbb{N}$. We have $|p(x)| < 1$ for all $x \geq M$. Moreover, for any $k \in \mathbb{N}$ the highest exponent in p^k is equal to kc_1, which implies that the dominant exponent of p^k tends to $-\infty$ as k tends to ∞. As a consequence, the coefficients and the exponents of more and more successive terms of the partial sum function s_n that is defined by $s_n(x) = \sum_{k=0}^{n} (-1)^k p^k(x)$ for $x \in [M, \infty)$ will not change any more as n becomes larger and larger. Therefore, the

series on the right-hand side of (44) also is a sum of exponentials:

$$\frac{1}{1+p(x)} = \sum_{k=0}^{\infty} (-1)^k \left(\sum_{i=1}^{\infty} \gamma_i e^{c_i x} \right)^k = \sum_{k=0}^{\infty} d_i e^{\delta_i x} \qquad \text{for all } x \geq M\,.$$

Note that the set of exponents of this series will have no finite accumulation point since the highest exponent in p^k is equal to kc_1. Let us now prove that this series also converges absolutely. Define $p^*(x) = \sum_{i=1}^{\infty} |\gamma_i| e^{c_i x}$ for all $x \geq P$. Since the terms of the series p^* are the absolute values of the terms of the series p and since p converges absolutely in $[P, \infty)$, p^* also converges absolutely in $[P, \infty)$. By Lemma A.1 the series p^* also converges uniformly in $[P, \infty)$. Furthermore, $\{c_i\}_{i=1}^{\infty}$ is a nonincreasing and unbounded sequence of negative numbers. As a consequence, we have $\lim_{x \to \infty} p^*(x) = 0$ (cf. (43)). So $|\, p^*(x)\,|$ will be less than one if x is large enough, say if $x \geq N$. Therefore, we have

$$\frac{1}{1+p^*(x)} = \sum_{k=0}^{\infty} (-1)^k \left(p^*(x)\right)^k \qquad \text{for all } x \geq N\,.$$

This series converges absolutely in $[N, \infty)$. Since

$$\sum_{k=0}^{\infty} |d_i| e^{\delta_i x} \leq \sum_{k=0}^{\infty} \left(\sum_{i=1}^{\infty} |\gamma_i| e^{c_i x} \right)^k = \sum_{k=0}^{\infty} \left| \left(p^*(x)\right)^k \right|,$$

the series $\sum_{k=0}^{\infty} d_i e^{\delta_i x}$ also converges absolutely for any $x \in [N, \infty)$. Since this series converges absolutely, we can reorder the terms. After reordering the terms, adding up terms with the same exponents and removing terms of the form $re^{\varepsilon x}$ with $r \in \mathbb{R}$ if necessary, the sequence of exponents will be decreasing and unbounded from below.

This implies that $\frac{1}{1+p} \in \mathcal{S}_{\mathrm{e}}$ and thus also $\frac{1}{f} \in \mathcal{S}_{\mathrm{e}}$.

As a consequence, $\frac{g}{f} = g\frac{1}{f}$ also belongs to $\mathcal{S}_{\mathrm{e}}$.

If there exists a real number Q such that $f(x) > 0$ for all $x \geq Q$, then the function $\sqrt{f}$ is defined and analytic in $[Q, \infty)$. We may assume without loss of generality that $Q \geq K$ and that the domain of definition of f is $[Q, \infty)$.

If $a_0 = \varepsilon$, then we have $\sqrt{f(x)} = 0$ for all $x \geq Q$ and thus $\sqrt{f} \in \mathcal{S}_{\mathrm{e}}$.

If $a_0 \neq \varepsilon$, then $\alpha_0 > 0$ and then we can rewrite $\sqrt{f(x)}$ as

$$\sqrt{f(x)} = \sqrt{\alpha_0}\ e^{\frac{1}{2}a_0 x} \sqrt{1+p(x)}\,.$$

Now we can use the Taylor series expansion of $\sqrt{1+x}$. This leads to

$$\sqrt{1+p(x)} = \sum_{k=0}^{\infty} \frac{\Gamma\left(\frac{3}{2}\right)}{\Gamma\left(\frac{3}{2}-k\right)\, k!}\ p^k(x) \qquad \text{if } |p(x)| < 1\,,$$

where Γ is the gamma function. If we apply the same reasoning as for $\frac{1}{1+p}$, we find that $\sqrt{1+p} \in \mathcal{S}_{\mathrm{e}}$ and thus also $\sqrt{f} \in \mathcal{S}_{\mathrm{e}}$. □

Acknowledgments. The authors want to thank Prof. G. J. Olsder and Dr. S. Gaubert for their valuable comments on [7], from which a large part of the material presented in this paper has been extracted.

The authors also want to thank the anonymous reviewers for their useful comments and remarks, and for pointing out an error in the original proof of Lemma A.1.

This research was sponsored by the Concerted Action Project (GOA) of the Flemish Community, entitled "Model-based Information Processing Systems," by the Belgian program on interuniversity attraction poles (IUAP P4-02 and IUAP P4-24), and by the TMR Project "Algebraic Approach to Performance Evaluation of Discrete Event Systems (ALAPEDES)" of the European Commission.

Bart De Schutter is a senior research assistant with the F.W.O. (Fund for Scientific Research–Flanders). Bart De Moor is a research associate with the F.W.O.

REFERENCES

[1] F. Baccelli, G. Cohen, G. Olsder, and J. Quadrat, *Synchronization and Linearity*, John Wiley, New York, 1992.

[2] Z. Bai, *Note on the quadratic convergence of Kogbetliantz's algorithm for computing the singular value decomposition*, Linear Algebra Appl., 104 (1988), pp. 131–140.

[3] J.-P. Charlier and P. Van Dooren, *On Kogbetliantz's SVD algorithm in the presence of clusters*, Linear Algebra Appl., 95 (1987), pp. 135–160.

[4] J.-P. Charlier, M. Vanbegin, and P. Van Dooren, *On efficient implementations of Kogbetliantz's algorithm for computing the singular value decomposition*, Numer. Math., 52 (1988), pp. 279–300.

[5] R. Cuninghame-Green, *Minimax Algebra*, Lecture Notes in Econom. Math. Systems 166, Springer-Verlag, Berlin, 1979.

[6] R. Cuninghame-Green, *Using fields for semiring computations*, Ann. Discrete Math., 19 (1984), pp. 55–73.

[7] B. De Schutter, *Max-Algebraic System Theory for Discrete Event Systems*, Ph.D. thesis, Faculty of Applied Sciences, K.U. Leuven, Leuven, Belgium, 1996.

[8] B. De Schutter and B. De Moor, *The extended linear complementarity problem*, Math. Programming, 71 (1995), pp. 289–325.

[9] B. De Schutter and B. De Moor, *A method to find all solutions of a system of multivariate polynomial equalities and inequalities in the max algebra*, Discrete Event Dynamic Systems: Theory and Applications, 6 (1996), pp. 115–138.

[10] B. De Schutter and B. De Moor, *The singular value decomposition in the extended max algebra*, Linear Algebra Appl. 250 (1997), pp. 143–176.

[11] K. Fernando, *Linear convergence of the row cyclic Jacobi and Kogbetliantz methods*, Numer. Math., 56 (1989), pp. 73–91.

[12] S. Gaubert, *Théorie des Systèmes Linéaires dans les Dioïdes*, Ph.D. thesis, Ecole Nationale Supérieure des Mines de Paris, France, 1992.

[13] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.

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

[15] V. Hari, *On the quadratic convergence of the serial singular value decomposition Jacobi methods for triangular matrices*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1076–1096.

[16] V. Hari, *On sharp quadratic convergence bounds for the serial Jacobi methods*, Numer. Math., 60 (1991), pp. 375–406.

[17] V. Hari and K. Veselić, *On Jacobi methods for singular value decompositions*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 741–754.

[18] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[19] J. Hyslop, *Infinite Series*, 5th ed., Oliver and Boyd, Edinburgh, UK, 1954.

[20] E. Kogbetliantz, *Solution of linear equations by diagonalization of coefficients matrix*, Quart. Appl. Math., 13 (1955), pp. 123–132.

[21] S. Kung, *A new identification and model reduction algorithm via singular value decomposition*, in Proc. 12th Asilomar Conference on Circuits, Systems and Computers, Pacific Grove, CA, 1978, pp. 705–714.

[22] L. Ljung, *System Identification: Theory for the User*, Prentice–Hall, Englewood Cliffs, NJ, 1987.

[23] S. Mandelbrojt, *Dirichlet Series*, D. Reidel, Dordrecht, The Netherlands, 1972.

[24] J. Marsden, *Elementary Classical Analysis*, W.H. Freeman, San Francisco, CA, 1974.

[25] Max Plus, *Linear systems in (max,+) algebra*, in Proc. 29th IEEE Conference on Decision and Control, Honolulu, HI, 1990, pp. 151–156.

[26] G. Olsder, *Some Results on the Minimal Realization of Discrete-Event Dynamic Systems*, Tech. report 85-35, Delft University of Technology, Faculty of Technical Mathematics and Informatics, Delft, the Netherlands, 1985.

[27] G. Olsder, *On the characteristic equation and minimal realizations for discrete-event dynamic systems*, in Proc. 7th Internat. Conference on Analysis and Optimization of Systems, Antibes, France, Lecture Notes in Control and Inform. Sci. 83, Springer-Verlag, Berlin, 1986, pp. 189–201.

[28] G. Olsder and R. de Vries, *On an analogy of minimal realizations in conventional and discrete-event dynamic systems*, in Algèbres Exotiques et Systèmes à Evénements Discrets, CNRS/CNET/INRIA Seminar, CNET, Issy-les-Moulineaux, France, 1987, pp. 193–213.

[29] G. Olsder and C. Roos, *Cramér and Cayley-Hamilton in the Max-Algebra*, Tech. report 85-30, Delft University of Technology, Faculty of Technical Mathematics and Informatics, Delft, The Netherlands, 1985.

[30] G. Olsder and C. Roos, *Cramer and Cayley-Hamilton in the max algebra*, Linear Algebra App., 101 (1988), pp. 87–108.

[31] C. Paige and P. Van Dooren, *On the quadratic convergence of Kogbetliantz's algorithm for computing the singular value decomposition*, Linear Algebra Appl., 77 (1986), pp. 301–313.

[32] G. Strang, *Linear Algebra and Its Applications*, 3rd ed., Harcourt, Brace, and Jovanovich, Fort Worth, TX, 1988.

[33] P. Van Overschee and B. De Moor, *Subspace algorithms for the stochastic identification problem*, Automatica, 29 (1993), pp. 649–660.

[34] P. Van Overschee and B. De Moor, *N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems*, Automatica, 30 (1994), pp. 75–93.

[35] M. Verhaegen, *Identification of the deterministic part of MIMO state space models given in innovations form from input-output data*, Automatica, 30 (1994), pp. 61–74.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 407–415, April 1998

008

ON THE BLOCK INDEPENDENCE IN REFLEXIVE INNER INVERSE AND M–P INVERSE OF BLOCK MATRIX*

YIJU WANG†

Abstract. We examine the reflexive inner inverse and M–P inverse of block matrix. First, we give the definitions of block independence in generalized inverse of block matrix, and derive necessary and sufficient conditions for two $m \times n$ matrices being block independent in reflexive inner inverse and in M–P inverse. An analogous set of conditions for three ordered $m \times n$ matrices is also derived in this paper.

Key words. reflexive inner inverse, M–P inverse, block matrix, block independence, rank of matrix

AMS subject classifications. 15A04, 15A09, 15A18

PII. S0895479896301868

1. Introduction.

Let $A \in C^{m\times n}$ and consider the following four Moore–Penrose equations:

$$AGA = A, \tag{1}$$

$$GAG = G, \tag{2}$$

$$(AG)^* = AG, \tag{3}$$

$$(GA)^* = GA. \tag{4}$$

Suppose $\mathcal{J} = \{i, j, ..., k\}$ is a nonempty subset of $\{1, 2, 3, 4\}$; then a matrix G is said to be a $\mathcal{J}$-inverse of A if G satisfies equation (i) for each $i \in \mathcal{J}$. The set of all $\mathcal{J}$-inverse of A is denoted by $A^{\{\mathcal{J}\}}$ and its any element is denoted by $A^{\mathcal{J}}$. $\{1\}$-*inverse*, $\{1, 2\}$-*inverse*, and $\{1, 2, 3, 4\}$-*inverse* are also called inner inverse, reflexive inner inverse, and M–P (Moore–Penrose) inverse of A, denoted by A^-, A^G, and A^+, respectively.

Throughout this paper, all our matrices will be over the complex number field C. For a matrix A in the set $C^{m\times n}$, the symbols A^*, $\mathrm{rk}(A)$, $\mathcal{R}(A)$, $\mathcal{N}(A)$, $\mathcal{RS}(A)$, and $\mathrm{tr}(A)$ denote the conjugate transpose, the rank, the range (column space), the nullspace, the row space, and the trace of A, respectively. In the following, we suppose $\mathcal{J} = \{i, j, ..., k\}$ is a nonempty subset of $\{1, 2, 3, 4\}$.

DEFINITION 1.1. *We say two $m \times n$ matrices B and C are block independent in $\mathcal{J}$-inverse if there exist $B^{\mathcal{J}} \in B^{\{\mathcal{J}\}}, C^{\mathcal{J}} \in C^{\{\mathcal{J}\}}$ such that*

$$\begin{pmatrix} B^{\mathcal{J}} & C^{\mathcal{J}} \end{pmatrix} \in \begin{pmatrix} B \\ C \end{pmatrix}^{\{\mathcal{J}\}} \quad \text{and} \quad \begin{pmatrix} B^{\mathcal{J}} \\ C^{\mathcal{J}} \end{pmatrix} \in \begin{pmatrix} B & C \end{pmatrix}^{\{\mathcal{J}\}}.$$

*Received by the editors April 10, 1996; accepted for publication (in revised form) by G. P. Styan March 17, 1997. This research was partially supported by NSF of Shandong Province, China.
http://www.siam.org/journals/simax/19-2/30186.html
†Institute of Operations Research, Qufu Normal University, Qufu Shangdong Province, P.R. China 273165.

DEFINITION 1.2. *We say three ordered $m \times n$ matrices A, B, C are block independent in $\mathcal{J}$-inverse if there exist $A^{\mathcal{J}} \in A^{\{\mathcal{J}\}}, B^{\mathcal{J}} \in B^{\{\mathcal{J}\}}, C^{\mathcal{J}} \in C^{\{\mathcal{J}\}}$ such that*

$$\begin{pmatrix} A^{\mathcal{J}} & C^{\mathcal{J}} \\ B^{\mathcal{J}} & 0 \end{pmatrix} \in \begin{pmatrix} A & B \\ C & 0 \end{pmatrix}^{\{\mathcal{J}\}}.$$

DEFINITION 1.3. *We say four ordered $m \times n$ matrices A, B, C, and D are block independent in $\mathcal{J}$-inverse if there exist $A^{\mathcal{J}} \in A^{\{\mathcal{J}\}}, B^{\mathcal{J}} \in B^{\{\mathcal{J}\}}, C^{\mathcal{J}} \in C^{\{\mathcal{J}\}}, D^{\mathcal{J}} \in D^{\{\mathcal{J}\}}$ such that*

$$\begin{pmatrix} A^{\mathcal{J}} & C^{\mathcal{J}} \\ B^{\mathcal{J}} & D^{\mathcal{J}} \end{pmatrix} \in \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{\{\mathcal{J}\}}.$$

It should be noted that the definition of independence of blocks of generalized inverse of block matrix in another meaning have been given in [1, 2] by Hall and in [3] by Hall and Hartwig. But for clarity of the difference between these two kinds of definitions we repeat their definition.

DEFINITION 1.4. *Let*

$$G_1 = \begin{pmatrix} Q_1 \\ L_1 \end{pmatrix} \quad \text{and} \quad G_2 = \begin{pmatrix} Q_2 \\ L_2 \end{pmatrix}$$

be two possibly different $\{1\}$-inverses of $(A \;\; B)$. Then the blocks of all $\{1\}$-inverses of $(A \;\; B)$ are said to be independent whenever $G = \begin{pmatrix} Q_1 \\ L_2 \end{pmatrix}$ is a $\{1\}$-inverse of $(A \;\; B)$ for every possible choice of G_1 and G_2.

Independence of blocks of $\{1,3\}$- and $\{1,4\}$-inverses of $(A \;\; B)$ and of generalized inverses of $\begin{pmatrix} A \\ C \end{pmatrix}$ is defined similarly.

DEFINITION 1.5. *For general block matrix*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

let G_1, G_2, G_3, and G_4 be four possibly different $\{1\}$-inverses of M, and let $G_i(jk)$ denote the (j,k) block of G_i, where $i = 1, 2, 3$, or 4, while $j, k = 1$ or 2. Then the blocks of all $\{1\}$-inverses of M are said to be independent whenever

$$G = \begin{pmatrix} G_1(11) & G_2(12) \\ G_3(21) & G_4(22) \end{pmatrix}$$

is a $\{1\}$-inverse of M for every possible choice of G_1, G_2, G_3, and G_4.

The independence of blocks of $\{1,3\}$- and $\{1,4\}$-inverses of M is defined similarly.

It is easy to see that these two kinds of definitions are different. First, all blocks in $\mathcal{J}$-inverse of block matrix in Definitions 1.2 and 1.3 are requested to be of the same order but not in the definition by Hall and Hartwig. Second, for any nonempty $\mathcal{J}' \subset \mathcal{J} \subseteq \{1,2,3,4\}$, if two $m \times n$ complex matrices A, B are independent in $\mathcal{J}$-inverse, then they are independent in $\mathcal{J}'$-inverse, but for the block matrix M in Definition 1.5, the blocks of $\{1\}$-inverses of M are independent of each other if and only if the blocks of both $\{1,3\}$- and $\{1,4\}$-inverses of M are independent (cf. Theorems 1, 2, and 3 in [2]). The relations between these two kinds of definitions are discussed at the end of the following sections.

In this paper, we mainly examine the block independence in reflexive inner inverse and in M–P inverse. The outline of our paper is as follows. In section 2 we examine the block independence in reflexive inner inverse of two $m \times n$ complex matrices as well as in M–P inverse, and derive necessary and sufficient conditions for two $m \times n$ matrices being block independent in reflexive inner inverse as well as in M–P inverse. In section 3, we examine the block independence in reflexive inner inverse of three ordered $m \times n$ complex matrices as well as in M–P inverse. In particular, necessary and sufficient conditions for three ordered $m \times n$ matrices being block independent in reflexive inner inverse as well as in M–P inverse are also derived in this section, which generalizes Theorem 2.1 of Markham and Fiedler [8]. And in the end of each section, we investigate the relations between two kinds of definitions of independence given in this paper.

2. Block independence of two $m \times n$ complex matrices. First, we give the concepts of $*$-*orthogonality* and $*$-*commutativity* which were introduced by Hestenes [5].

DEFINITION 2.1. *Two $m \times n$ complex matrices A and B are said to be $*$-orthogonal if $A^*B = 0$ and $AB^* = 0$. Two $m \times n$ complex matrices A and B are said to be $*$-commute if $A^*B = B^*A$ and $AB^* = BA^*$. Moreover, we say that A and B are rank additive (or their rank is additive) if* $\mathrm{rk}(A + B) = \mathrm{rk}(A) + \mathrm{rk}(B)$.

The connection between $*$-*orthogonality*, $*$-*commutativity*, and rank additivity can be seen in the following lemma.

LEMMA 2.2. *Two $m \times n$ complex matrices A and B are $*$-orthogonal if and only if they are $*$-commute and their rank is additive.*

Proof. Only if: Trivially $*$-*orthogonality* implies $*$-*commutativity*. We have $A^*B = 0 \Leftrightarrow \mathcal{R}(B) \subseteq \mathcal{N}(A^*) \Rightarrow \mathcal{R}(B) \cap \mathcal{R}(A) = \{0\}$ as well as $AB^* = 0 \Leftrightarrow \mathcal{R}(B^*) \subseteq \mathcal{N}(A) \Rightarrow \mathcal{R}(B^*) \cap \mathcal{R}(A^*) = \{0\}$. But $\mathcal{R}(B) \cap \mathcal{R}(A) = \{0\}$ and $\mathcal{R}(B^*) \cap \mathcal{R}(A^*) = \{0\}$ are equivalent to $\mathrm{rk}(A + B) = \mathrm{rk}(A) + \mathrm{rk}(B)$ from Theorem 11 in Marsaglia and Styan [6].

If: since $AB^* = BA^*$ and $\mathcal{R}(A) \cap \mathcal{R}(B) = \{0\}$, one has $AB^* = 0$. Similarly, we have $A^*B = 0$ since A^*B is a Hermitian matrix and $\mathcal{R}(A^*) \cap \mathcal{R}(B^*) = \{0\}$. ∎

It is clear that $A^*B = 0 \Leftrightarrow A^+B = 0 \Leftrightarrow B^+A = 0$ as well as $AB^* = 0 \Leftrightarrow AB^+ = 0 \Leftrightarrow BA^+ = 0$. Hence, we have the following.

THEOREM 2.3. *Two $m \times n$ complex matrices are block independent in reflexive inner inverse if and only if their rank is additive. Moreover they are block independent in M–P inverse if and only if they are $*$-orthogonal.*

Proof. We only give the proof of the first part of the theorem since the latter part can easily be seen.

Let B, C be two $m \times n$ matrices if there exist $\{1,2\}$-inverses B^G, C^G of B and C, respectively, such that

$$B^GC = 0, \qquad BC^G = 0, \qquad C^GB = 0, \qquad CB^G = 0.$$

It can easily be verified that

$$\begin{pmatrix} B^G & C^G \end{pmatrix} \in \begin{pmatrix} B \\ C \end{pmatrix}^{\{1,2\}}, \qquad \begin{pmatrix} B^G \\ C^G \end{pmatrix} \in \begin{pmatrix} B & C \end{pmatrix}^{\{1,2\}}.$$

Hence, B, C are block independent in reflexive inner inverse.

Conversely, by

$$\begin{pmatrix} B^G & C^G \end{pmatrix} \in \begin{pmatrix} B \\ C \end{pmatrix}^{\{1,2\}},$$

one has

$$\begin{pmatrix} B^G & C^G \end{pmatrix} \begin{pmatrix} B \\ C \end{pmatrix} \begin{pmatrix} B^G & C^G \end{pmatrix} = \begin{pmatrix} B^G & C^G \end{pmatrix},$$

so

$$B^G B C^G = 0, \quad C^G C B^G = 0.$$

Multiplying the left side by B, C, respectively, one has

$$BC^G = 0, \qquad CB^G = 0.$$

Similarly, by

$$\begin{pmatrix} B^G \\ C^G \end{pmatrix} \in \begin{pmatrix} B & C \end{pmatrix}^{\{1,2\}},$$

one has $B^G C = 0$, $C^G B = 0$. □

Let A, B be two $m \times n$ complex matrices; denote $W_1 = \begin{pmatrix} A & B \end{pmatrix}, W_2 = \begin{pmatrix} A \\ B \end{pmatrix}$, respectively. From the above theorem, Theorem 11 in Marsaglia and Styan [6], and Theorems 2.1 and 2.2 in [3], we know that

A and B are block independent in reflexive inner inverse

$\Leftrightarrow \operatorname{rk}(A+B) = \operatorname{rk}(A) + \operatorname{rk}(B)$

$\Leftrightarrow \mathcal{R}(A) \cap \mathcal{R}(B) = \{0\}$ and $\mathcal{R}(A^*) \cap \mathcal{R}(B^*) = \{0\}$

$\Leftrightarrow {W_1}^+ W_1$ and $W_2 {W_2}^+$ are both block diagonal

$\Leftrightarrow$ the blocks in the $\{1\}$-*inverses* for W_1 and W_2 are independent of each other.

Hence, we obtain a relationship between two kinds of definitions of independence for block matrix which has two blocks.

PROPERTY 1. *Two $m \times n$ complex matrices A and B are block independent in reflexive inner inverse if and only if the blocks of $\{1\}$-inverses for $\begin{pmatrix} A & B \end{pmatrix}$ and $\begin{pmatrix} A \\ B \end{pmatrix}$ are independent of each other.*

3. Block independence of three $m \times n$ complex matrices. In this section, we denote

$$M_1 := \begin{pmatrix} A & B \\ C & 0 \end{pmatrix}, \qquad M_2 := \begin{pmatrix} 0 & B \\ C & D \end{pmatrix},$$

$$M_3 := \begin{pmatrix} A & B \\ 0 & D \end{pmatrix}, \qquad M_4 := \begin{pmatrix} A & 0 \\ C & D \end{pmatrix},$$

$$M := \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

DEFINITION 3.1. *Let A, B, C be $m \times n$ complex matrices, and let A^G, B^G, C^G be given respective $\{1,2\}$-inverses of A, B, C. Then the ordered pair (B^G, C^G) is said to be related via A^G if*

$$B^G A = 0, \qquad A^G B = 0, \qquad C^G A = 0, \qquad A C^G = 0$$

hold.

Note that for given A^G and B^G, the pair (B^G, B^G) is related via A^G if and only if B^G and A^G are related in the sense of Fiedler and Markham [8, p. 167]. Hence, the above definition can be seen as a generalization of the relation of two $\{1,2\}$-inverses to the relation of three $\{1,2\}$-inverses. Note that the relation of pair (B^G, C^G) via A^G is not the same as the relation of the relation of the pair (C^G, B^G) via A^G, as will become clearer subsequently.

THEOREM 3.2. *Let A, B, C be three $m \times n$ complex matrices. Then the following statements are equivalent:*

(1) *three ordered matrices A, B, C are block independent in reflexive inner inverse;*

(2) *there exist $\{1,2\}$-inverses A^G, B^G, and C^G of A, B, and C, respectively, such that the ordered pair (B^G, C^G) is related via A^G;*

(3) $\mathrm{rk}(M_1) = \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C)$;

(4) $\mathcal{R}(A) \cap \mathcal{R}(B) = \{0\}$ *and* $\mathcal{R}(A^*) \cap \mathcal{R}(C^*) = \{0\}$.

Proof. The equivalence of (1) and (2) can easily be seen.

(3)$\Longrightarrow$(1): For $A, B, C \in C^{m\times n}$, there exist nonsingular matrices P_1, Q_1 (cf. Theorem 6.2.2 in [7]) such that

$$P_1AQ_1 = \begin{pmatrix} I_A & 0 \\ 0 & 0 \end{pmatrix}, \qquad P_1BQ_1 = \begin{pmatrix} B_1 & B_2 \\ B_3 & B_4 \end{pmatrix}, \qquad P_1CQ_1 = \begin{pmatrix} C_1 & C_2 \\ C_3 & C_4 \end{pmatrix}.$$

Since $\mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C) = \mathrm{rk}(M_1) \le \mathrm{rk}(A \;\; B) + \mathrm{rk}(C) \le \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C)$, we have $\mathrm{rk}(A \;\; B) = \mathrm{rk}(A) + \mathrm{rk}(B)$ and $\mathcal{R}(A) \cap \mathcal{R}(B) = \{0\}$ (cf. Theorem 11 in [6]). By

$$\begin{aligned} \mathrm{rk}(M_1) &= \mathrm{rk}\left[\begin{pmatrix} P_1 & 0 \\ 0 & P_1 \end{pmatrix} \begin{pmatrix} A & B \\ C & 0 \end{pmatrix} \begin{pmatrix} Q_1 & 0 \\ 0 & Q_1 \end{pmatrix} \right] \\ &= \mathrm{rk} \begin{pmatrix} I_A & 0 & B_1 & B_2 \\ 0 & 0 & B_3 & B_4 \\ C_1 & C_2 & 0 & 0 \\ C_3 & C_4 & 0 & 0 \end{pmatrix} \\ &= \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C), \end{aligned}$$

and $\mathcal{R}(A) \cap \mathcal{R}(B) = \{0\}$, we have $\mathrm{rk}(B_3 \;\; B_4) = \mathrm{rk}(B)$ and there exists nonsingular matrix P_2 of the form $P_2 = \begin{pmatrix} I_A & * \\ 0 & I \end{pmatrix}$ such that

$$P_2P_1AQ_1 = \begin{pmatrix} I_A & 0 \\ 0 & 0 \end{pmatrix}, \qquad P_2P_1BQ_1 = \begin{pmatrix} 0 & 0 \\ B_3 & B_4 \end{pmatrix}, \qquad P_2P_1CQ_1 = \begin{pmatrix} \overline{C_1} & \overline{C_2} \\ C_3 & C_4 \end{pmatrix}.$$

Hence,

$$\begin{aligned} \mathrm{rk}(M_1) &= \mathrm{rk}\left[\begin{pmatrix} P_2P_1 & 0 \\ 0 & P_2P_1 \end{pmatrix} \begin{pmatrix} A & B \\ C & 0 \end{pmatrix} \begin{pmatrix} Q_1 & 0 \\ 0 & Q_1 \end{pmatrix} \right] \\ &= \mathrm{rk} \begin{pmatrix} I_A & 0 & 0 & 0 \\ 0 & 0 & B_3 & B_4 \\ \overline{C_1} & \overline{C_2} & 0 & 0 \\ C_3 & C_4 & 0 & 0 \end{pmatrix} \\ &= \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C), \end{aligned}$$

and

$$\mathrm{rk} \begin{pmatrix} \overline{C_2} \\ C_4 \end{pmatrix} = \mathrm{rk}(C).$$

There exists nonsingular matrix Q_2 of the form $Q_2 = \begin{pmatrix} I_A & 0 \\ * & I \end{pmatrix}$ such that

$$P_2P_1AQ_1Q_2 = \begin{pmatrix} I_A & 0 \\ 0 & 0 \end{pmatrix}, \quad P_2P_1BQ_1Q_2 = \begin{pmatrix} 0 & 0 \\ \overline{B_3} & B_4 \end{pmatrix}, \quad P_2P_1CQ_1Q_2 = \begin{pmatrix} 0 & \overline{C_2} \\ 0 & C_4 \end{pmatrix}.$$

Let $P = P_2P_1, Q = Q_1Q_2$, respectively; one has

$$PAQ = \begin{pmatrix} I_A & 0 \\ 0 & 0 \end{pmatrix}, \qquad PBQ = \begin{pmatrix} 0 & 0 \\ \overline{B_3} & B_4 \end{pmatrix}, \qquad PCQ = \begin{pmatrix} 0 & \overline{C_2} \\ 0 & C_4 \end{pmatrix}.$$

Let

$$A^G = Q \begin{pmatrix} I_A & 0 \\ 0 & 0 \end{pmatrix} P, \qquad B^G = Q \begin{pmatrix} 0 & \widehat{B_3} \\ 0 & \widehat{B_4} \end{pmatrix} P, \qquad C^G = Q \begin{pmatrix} 0 & 0 \\ \widehat{C_2} & \widehat{C_4} \end{pmatrix} P,$$

where $\begin{pmatrix} \widehat{B_3} \\ \widehat{B_4} \end{pmatrix}$ is a reflexive inner inverse of $(\overline{B_3} \;\; B_4)$, $(\widehat{C_2} \;\; \widehat{C_4})$ is a reflexive inner inverse of $\begin{pmatrix} \overline{C_2} \\ C_4 \end{pmatrix}$. It can easily be verified that

$$AC^G = 0, \quad CA^G = 0, \quad B^GA = 0, \quad A^GB = 0,$$

and

$$\widehat{M_1} := \begin{pmatrix} A^G & C^G \\ B^G & 0 \end{pmatrix}$$

is a reflexive inner inverse of M_1.

(1) $\Longrightarrow$ (3): If there exist $A^G \in A^{\{1,2\}}, B^G \in B^{\{1,2\}}, C^G \in C^{\{1,2\}}$ such that

$$\widehat{M_1} := \begin{pmatrix} A^G & C^G \\ B^G & 0 \end{pmatrix} \in \begin{pmatrix} A & B \\ C & 0 \end{pmatrix}^{\{1,2\}},$$

by $M_1{M_1}^GM_1 = M_1$ and ${M_1}^GM_1{M_1}^G = {M_1}^G$, we have

$$AC^G = 0, \quad CA^G = 0, \quad B^GA = 0, \quad A^GB = 0,$$

and

$$M_1\widehat{M_1} = \begin{pmatrix} AA^G + BB^G & 0 \\ 0 & CC^G \end{pmatrix}$$

is a projector. Hence,

$$\begin{aligned} \mathrm{rk}(M_1) &= \mathrm{rk}(M_1\widehat{M_1}) \\ &= \mathrm{tr}(AA^G) + \mathrm{tr}(BB^G) + \mathrm{tr}(CC^G) \\ &= \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C). \end{aligned}$$

(3) $\Longleftrightarrow$ (4): It is clear that

$$\mathrm{rk}\begin{pmatrix} A^* & C^* \\ B^* & 0 \end{pmatrix} = \mathrm{rk}\begin{pmatrix} A^* \\ B^* \end{pmatrix} + \mathrm{rk}\begin{pmatrix} C^* \\ 0 \end{pmatrix} - \dim\left(\mathcal{R}\begin{pmatrix} A^* \\ B^* \end{pmatrix} \cap \mathcal{R}\begin{pmatrix} C^* \\ 0 \end{pmatrix}\right),$$

where $\mathrm{rk}\begin{pmatrix} C^* \\ 0 \end{pmatrix} = \mathrm{rk}(C)$ and

$$\mathrm{rk}\begin{pmatrix} A^* \\ B^* \end{pmatrix} = \mathrm{rk}(A) + \mathrm{rk}(B) - \dim(\mathcal{R}(A) \cap \mathcal{R}(B)).$$

Hence, we have $\mathrm{rk}(M_1) = \mathrm{rk}(A)+\mathrm{rk}(B)+\mathrm{rk}(C)$ if and only if $\mathcal{R}(A)\cap\mathcal{R}(B) = \{0\}$ and $\mathcal{R}\begin{pmatrix} A^* \\ B^* \end{pmatrix}\cap\mathcal{R}\begin{pmatrix} C^* \\ 0 \end{pmatrix} = \{0\}$, the latter being equivalent to $\mathcal{R}[A^*(I-BB^+)]\cap\mathcal{R}(C^*) = \{0\}$. But in view of $\mathcal{R}(A) \cap \mathcal{R}(B) = \{0\}$ we have $\mathrm{rk}[A^*(I - BB^+)] = \mathrm{rk}(A^*)$ (cf. Theorem 5 in [6]), and hence $\mathcal{R}[A^*(I - BB^+)] = \mathcal{R}(A^*)$. □

From the above theorem it becomes immediately clear that the relation of a pair (B^G, C^G) via A^G is not the same as the relation of (C^G, B^G) via A^G. Moreover, it is seen that the equivalence of (i) and (v) in Theorem 2.1 of [8] follows as a corollary from the above results.

Similarly, we have the following.

THEOREM 3.3. *Let $A, B, C, D \in C^{m\times n}$; then*

(1) *there exist $\{1,2\}$-inverses B^G, C^G, D^G of B, C, D, respectively, such that*

$$\begin{pmatrix} 0 & C^G \\ B^G & D^G \end{pmatrix} \in {M_2}^{\{1,2\}} \quad \textit{if and only if} \quad \mathrm{rk}(M_2) = \mathrm{rk}(B) + \mathrm{rk}(C) + \mathrm{rk}(D);$$

(2) *there exist $\{1,2\}$-inverses A^G, B^G, D^G of A, B, D, respectively, such that*

$$\begin{pmatrix} A^G & 0 \\ B^G & D^G \end{pmatrix} \in {M_3}^{\{1,2\}} \quad \textit{if and only if} \quad \mathrm{rk}(M_3) = \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(D);$$

(3) *there exist $\{1,2\}$-inverses A^G, C^G, D^G of A, C, D, respectively, such that*

$$\begin{pmatrix} A^G & C^G \\ 0 & D^G \end{pmatrix} \in {M_4}^{\{1,2\}} \quad \textit{if and only if} \quad \mathrm{rk}(M_4) = \mathrm{rk}(A) + \mathrm{rk}(C) + \mathrm{rk}(D).$$

Before examining the block indepdence in M–P inverse, we give the following lemma without proving.

LEMMA 3.4. *Let A, B, C be three $m \times n$ complex matrices. Then $AC^* = 0$ and $A^*B = 0$ if and only if $A^*AC^*C = C^*CA^*A$ and $AA^*BB^* = BB^*AA^*$ hold together with $\mathcal{R}(A) \cap \mathcal{R}(B) = \{0\}$ and $\mathcal{R}(A^*) \cap \mathcal{R}(C^*) = \{0\}$.*

From this lemma, it is easy to prove the following.

THEOREM 3.5. *Let A, B, C be three $m \times n$ complex matrices. Then three ordered matrices A, B, and C are block independent in M–P inverse if and only if one of the following conditions holds:*

(1) *$AC^* = 0$ and $A^*B = 0$;*

(2) *$AC^+ = 0$ and $A^+B = 0$;*

(3) *$A^*AC^*C = C^*CA^*A$, $AA^*BB^* = BB^*AA^*$, $\mathcal{R}(A)\cap\mathcal{R}(B) = \{0\}$, $\mathcal{R}(A^*)\cap\mathcal{R}(C^*) = \{0\}$;*

(4) *$\mathrm{rk}(M_1) = \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C)$ and $A^*AC^*C = C^*CA^*A$, $AA^*BB^* = BB^*AA^*$.*

It is clear that the three ordered $m \times n$ matrices A, B, C are independent in M–P inverse if and only if they are independent in reflexive inner inverse together with $A^*AC^*C = C^*CA^*A$ and $AA^*BB^* = BB^*AA^*$ by Lemma 3.4 and Theorem 3.3. Similarly, we have the following.

THEOREM 3.6. *Let $A, B, C, D \in C^{m\times n}$; then*

(1) *${M_2}^+ = \begin{pmatrix} 0 & C^+ \\ B^+ & D^+ \end{pmatrix}$ if and only if $C^*D = 0$ and $BD^* = 0$;*

(2) ${M_3}^+ = \begin{pmatrix} A^+ & 0 \\ B^+ & D^+ \end{pmatrix}$ *if and only if* $A^*B = 0$ *and* $BD^* = 0$*;*

(3) ${M_4}^+ = \begin{pmatrix} A^+ & C^+ \\ 0 & D^+ \end{pmatrix}$ *if and only if* $AC^* = 0$ *and* $D^*C = 0$.

THEOREM 3.7. *If four ordered* $m \times n$ *matrices* A, B, C, D *are independent in reflexive inner inverse, then* $\mathrm{rk}(M) = \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C) + \mathrm{rk}(D)$.

Proof. By Definition 1.3, there exist $\{1,2\}$-inverses A^G, B^G, C^G, and D^G of A, B, C, and D, respectively, such that

$$\widehat{M} := \begin{pmatrix} A^G & C^G \\ B^G & D^G \end{pmatrix} \in M^{\{1,2\}} \quad \text{and} \quad M\widehat{M} = \begin{pmatrix} AA^G + BB^G & * \\ * & CC^G + DD^G \end{pmatrix}.$$

Since $M\widehat{M}$ is a projector, we have

$$\begin{aligned} \mathrm{rk}(M) &= \mathrm{rk}(M\widehat{M}) = \mathrm{tr}(M\widehat{M}) \\ &= \mathrm{tr}(AA^G + BB^G) + \mathrm{tr}(CC^G + DD^G) \\ &= \mathrm{tr}(AA^G) + \mathrm{tr}(BB^G) + \mathrm{tr}(CC^G) + \mathrm{tr}(DD^G) \\ &= \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C) + \mathrm{rk}(D). \qquad \square \end{aligned}$$

Since

$$\begin{aligned} &\mathrm{rk}(M) \le \mathrm{rk}\begin{pmatrix} A \\ C \end{pmatrix} + \mathrm{rk}\begin{pmatrix} B \\ D \end{pmatrix} \le \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C) + \mathrm{rk}(D), \\ &\mathrm{rk}(M) \le \mathrm{rk}(A \;\; B) + \mathrm{rk}(C \;\; D) \le \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C) + \mathrm{rk}(D), \end{aligned}$$

and from the above theorem, Theorem 1 in [2], and Theorems 2.1 and 2.2 in [3], we know that four ordered $m \times n$ complex matrices A, B, C, and D are independent in reflexive inner inverse

$$\begin{aligned} &\Rightarrow \mathrm{rk}(M) = \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C) + \mathrm{rk}(D), \\ &\Rightarrow \mathrm{rk}(M) = \mathrm{rk}(A \;\; B) + \mathrm{rk}(C \;\; D) = \mathrm{rk}\begin{pmatrix} A \\ C \end{pmatrix} + \mathrm{rk}\begin{pmatrix} B \\ D \end{pmatrix}, \\ &\Leftrightarrow \mathcal{R}\begin{pmatrix} A \\ C \end{pmatrix} \cap \mathcal{R}\begin{pmatrix} B \\ D \end{pmatrix} = \{0\} \text{ and } \mathcal{R}\begin{pmatrix} A^* \\ B^* \end{pmatrix} \cap \mathcal{R}\begin{pmatrix} C^* \\ D^* \end{pmatrix} = \{0\}, \\ &\Leftrightarrow M^+M \text{ and } MM^+ \text{ are both block diagonal}, \\ &\Leftrightarrow \text{the blocks in the } \{1\}\text{-inverses for } M \text{ are independent of each other.} \end{aligned}$$

Hence, we also obtain a relationship between two kinds of definitions of independence of block matrix M.

PROPERTY 2. *If four ordered* $m \times n$ *complex matrices A,B,C, and D are independent in reflexive inner inverse, then the blocks in the* $\{1\}$*-inverses for M are independent of each other. In particular, if three ordered* $m \times n$ *complex matrices A,B, and C are independent in reflexive inner inverse, then the blocks in the* $\{1\}$*-inverses for* M_1 *are independent of each other.*

Now we have derived the necessary and sufficient conditions for three ordered $m \times n$ matrices being independent in reflexive inner inverse as well as in M–P inverse, but the necessary and sufficient conditions for four ordered $m \times n$ matrices being independent in $\mathcal{J}$-inverse such as $\mathcal{J} = \{1,2\}$ as well as $\mathcal{J} = \{1,2,3,4\}$ are not derived in this paper. In the end of this paper, we pose the following conjecture.

CONJECTURE. *Four ordered* $m \times n$ *matrices* A, B, C, D *are independent in reflexive inner inverse if and only if* $\mathrm{rk}(M) = \mathrm{rk}(A) + \mathrm{rk}(B) + \mathrm{rk}(C) + \mathrm{rk}(D)$.

Acknowledgment. The author wishes to thank the referee for his helpful suggestions and valuable comments.

REFERENCES

[1] F. J. Hall, *Generalized inverses of a bordered matrix of operators*, SIAM J. Appl. Math., 29 (1975), pp. 152–163.
[2] F. J. Hall, *On the independence of blocks of generalized inverses of bordered matrices*, Linear Algebra Appl., 14 (1976), pp. 53–61.
[3] F. J. Hall and R. E. Hartwig, *Further result on generalized inverses of partitioned matrices*, SIAM J. Appl. Math., 30 (1976), pp. 617–624.
[4] F. J. Hall and C. D. Meyer, Jr., *Generalized inverses of the fundamental bordered matrix used in linear estimation*, Sankhyā Ser. A, 37 (1975), pp. 428–438.
[5] M. R. Hestenes, *Relative hermitian matrices*, Pacific J. Math., 11 (1961), pp. 225–245.
[6] G. Marsaglia and G. P. H. Styan, *Equalities and inequalities for ranks of matrices*, Linear and Multilinear Algebra, 2 (1974), pp. 269–292.
[7] L. Mirsky, *An Introduction to Linear Algebra*, Clarendon Press, Oxford, 1955; reprinted Dover, New York, 1990.
[8] M. Fiedler and T. L. Markham, *Quasidirect addition of matrices and generalized inverses*, Linear Algebra Appl., 191 (1993), pp. 165–182.
[9] X. C. He and W. Y. Shun, *Introduction to Generalized Inverse of Matrices*, Jiangshu Science Press, Nanjing, China, 1990 (in Chinese).

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 416–428, April 1998

009

ON THE SANDWICH SEMIGROUPS OF GROUP BOOLEAN MATRICES*

CEN JIANMIAO†

Abstract. Let GM_n be the semigroup of all the $n \times n$ ($n \geq 2$) group Boolean matrices, and let R be a nonzero element in GM_n, where G is an n-order Abelian group. The sandwich semigroup of GM_n with the sandwich element R is denoted by $GM_n(R)$. The purpose of this paper is to discuss the Green's classes, idempotent elements, maximal subgroups, and regular elements in $GM_n(R)$. If G is an n-order cyclic group, our results are exactly the results of Wenchao Huang in [*Linear Algebra Appl.*, 25 (1979), pp. 135–160].

Key words. group Boolean matrix, sandwich semigroup, Green's class

AMS subject classification. 15A23

PII. S0895479895290383

1. Introduction and preliminaries. Let $B = \{0, 1\}$ be a Boolean algebra, and let M_n be the set of all the $n \times n$ ($n \geq 2$) matrices over B. In [1], Wenchao Huang discussed the sandwich semigroups of circulant Boolean matrices. In this paper, we will study the sandwich semigroups of the following group Boolean matrices.

Let G be an n-order group, $N = \{1, 2, \ldots, n\}$, and let φ: $N \to G$ be a one-to-one correspondence. For $A = (a_{ij}) \in M_n$, we say that A is a (G, φ)-(Boolean) matrix if there exists a map f_A from G to B such that

$$a_{ij} = f_A(\varphi(i)^{-1}\varphi(j)), \qquad i, \ \ j \in N. \tag{1.1}$$

This definition is due to Kai Wang (see [2]). (G, φ)-matrix is a generalized circulant matrix in [2]. Let GM_n^{φ} denote the set of all (G, φ)-matrices over B. For A, $D \in GM_n^{\varphi}$, and $a \in B$, define

$$f_{aA}(g) = af_A(g), \qquad f_{A+D}(g) = f_A(g) + f_D(g),$$

$$f_{AD}(g) = \sum_{l=1}^{n} f_A(\varphi(l))f_D(\varphi(l)^{-1}g), \qquad g \in G. \tag{1.2}$$

By (1.1), it is clear that aA, $A + D$, $AD \in GM_n^{\varphi}$. Hence, GM_n^{φ} is a semiring. Let Z, E, and H denote the zero matrix, unit matrix, and universal matrix in M_n, respectively. Obviously, Z, E, $H \in GM_n^{\varphi}$ for any φ. Write $E = (e_1, e_2, \ldots, e_n)$, where $e_1, e_2, \ldots, e_n$ are unit column vectors. Let ψ be another one-to-one correspondence from N to G. Write $P(\tau) = (e_{\tau(1)}, e_{\tau(2)}, \ldots, e_{\tau(n)})$, where $\tau = \psi^{-1}\varphi$. Then $P(\tau)$ is an n-order permutation matrix and $P(\tau)^{-1} = P(\tau)^T = P(\tau^{-1})$, where $P(\tau)^T$ denotes the transpose of $P(\tau)$. Therefore, we have the following.

PROPOSITION 1.1. *$A \in GM_n^{\psi}$ if and only if $P(\tau)^{-1}AP(\tau) \in GM_n^{\varphi}$.*

Proof. Let $A = (a_{ij}) \in GM_n^{\psi}$. Write $P(\tau)^{-1}AP(\tau) = (b_{ij})$. It is easily known that $b_{ij} = a_{\tau(i)\tau(j)}, i, \ j \in N$. Then, $b_{ij} = f_A(\psi(\tau(i))^{-1}\psi(\tau(j))) = f_{P(\tau)^TAP(\tau)}(\varphi(i)^{-1}\varphi(j))$, where $f_{P(\tau)^TAP(\tau)} = f_A$, by (1.1) and (1.2). Hence, $P(\tau)^{-1}AP(\tau) \in GM_n^{\varphi}$. Inversely, if $P(\tau)^{-1}AP(\tau) \in GM_n^{\varphi}$, then $A = P(\tau^{-1})^{-1}P(\tau)^{-1}AP(\tau)P(\tau^{-1}) \in GM_n^{\psi}$. ∎

*Received by the editors June 3, 1995; accepted for publication (in revised form) by G. P. Styan April 2, 1997.

http://www.siam.org/journals/simax/19-2/29038.html

†Department of Mathematics, Ningbo Normal College, Ningbo 315211, Zhejiang, P. R. China.

By the above proposition, GM_n^{ψ} is isomorphic to GM_n^{φ}. Hence, we fix φ. And, briefly, GM_n^{φ} can be denoted as GM_n and (G,φ)-matrix can be called a group matrix or G-matrix. If G is an n-order cyclic group, GM_n is isomorphic to C_n where C_n denotes the set of all n-order circulant Boolean matrices. In general, GM_n is not isomorphic to C_n. In fact, if G is not a cyclic group of order n, then there exists not certainly a map φ from N to G such that GM_n^{φ} is equal to C_n (see [5]). Thus, all results in this paper are generalizations of results in [1]. For convenience, we need the following results also.

Let $f_1, f_2, \ldots, f_n$ be n maps from G to B and $P_1, P_2, \ldots, P_n$ be n-group matrices by the following, defined respectively:

$$f_r(g_s) = \begin{cases} 1 \text{ for } r = s, \\ 0 \text{ for } r \neq s, \end{cases} \qquad r, s \in N \tag{1.3}$$

and

$$P_r = (P_{ij}^{(r)}) = (f_r(g_i^{-1}g_j)) \in GM_n, \qquad r \in N, \tag{1.4}$$

where $g_i = \varphi(i)$, $i \in N$. Then $P_1, P_2, \ldots, P_n$ are permutation matrices of order n. Write $\mathbf{P} = \{P_1, P_2, \ldots, P_n\} = \{P_g \mid g \in G\}$, where $P_g = P_{\varphi^{-1}(g)}$ for $g \in G$. Then we have the following.

PROPOSITION 1.2. *$P_rP_s = P_t$ if and only if $g_rg_s = g_t$ for r, s, $t \in N$. Further, $\mathbf{P}$ is group isomorphic to G and*

$$P_gP_h = P_{gh}, \quad \text{for } g, \quad h \in G. \tag{1.5}$$

Proof. For r, $s \in N$, if $g_rg_s = g_t$, then $g_i^{-1}g_j = g_t$ if and only if $g_r^{-1}g_i^{-1}g_j = g_s$ for $i, j \in N$. By (1.2), (1.3), and (1.4), $P_rP_s = (\sum_{k=1}^{n} f_r(g_k)f_s(g_k^{-1}g_i^{-1}g_j)) = (f_s(g_r^{-1}g_i^{-1}g_j)) = (f_t(g_i^{-1}g_j)) = P_t$. Inversely, if $P_rP_s = P_t$, then $f_s(g_r^{-1}g_i^{-1}g_j) = f_t(g_i^{-1}g_j)$ for i, $j \in N$. Choose $g_i = g_r^{-1}$ and $g_j = g_s$. Then $f_t(g_rg_s) = f_s(g_s) = 1$. Hence, $g_rg_s = g_t$ by (1.3). It is easily known that $\mathbf{P}$ is group isomorphic to G and (1.5) holds. □

PROPOSITION 1.3. *For $A \in M_n$, $A \in GM_n$ if and only if A can be written as*

$$A = \sum_{i=1}^{n} a_iP_i, \qquad a_i \in B, \quad \text{for } i \in N \tag{1.6}$$

or

$$A = \sum_{g \in G} a_gP_g, \qquad a_g \in B, \quad \text{for } g \in G. \tag{1.7}$$

And every element in GM_n has a unique representation in form (1.6) *or* (1.7).

Proof. Since $P_i \in GM_n$ for every $i \in N$, $\sum_{i=1}^{n} a_iP_i \in GM_n$ by (1.2). Inversely, if $A \in GM_n$, then we have $f_A = a_1f_1 + a_2f_2 + \cdots + a_nf_n$, where $a_i = f_A(g_i)$ for $i \in N$. Thus, $A = a_1P_1 + a_2P_2 + \cdots + a_nP_n$. If $A = \sum_{i=1}^{n} b_iP_i$ also, then $\sum_{i=1}^{n} a_if_i(g_s^{-1}g_t) = \sum_{i=1}^{n} b_if_i(g_s^{-1}g_t)$, s, $t \in N$. Thus, by (1.3), $a_r = \sum_{i=1}^{n} a_if_i(g_r) = \sum_{i=1}^{n} b_if_i(g_r) = b_r, r \in N$. Therefore, the representation is unique. □

Therefore, by the above propositions, if G is an Abelian group, the set GM_n forms a commutative semigroup under the usual multiplication operation of Boolean matrices. In this paper, we only discuss this case. For an arbitrary but fixed element

$R \in GM_n$, we can define an operation $*$ in GM_n as follows: for arbitrary A, $B \in GM_n$, $A * B = ARB$, where ARB is the usual product of Boolean matrices. It can be easily proved that GM_n is also a commutative semigroup under the operation $*$. We denote this semigroup by $GM_n(R)$ and call it a sandwich semigroup of group Boolean matrices with sandwich matrix R. The purpose of this paper is to discuss Green's relations, idempotent elements, regular elements, and maximal subgroups for $GM_n(R)$ with $R \neq Z$. If G is a cyclic group of order n, then our results are exactly the results of Wenchao Huang in [1]. We need the following notation also.

Let $Z \neq A \in GM_n$; then there exists a subset S of G such that

$$A = \sum_{g \in S} P_g. \tag{1.8}$$

We write $S = SN(A)$. $SN(A)$ is said to be the set of support elements of A. Let $R \neq Z$ be a fixed element in GM_n. For convenience, assume $SN(R) = \{r_1, r_2, \ldots, r_l\}$ throughout this paper. Let S be a subset of G. We denote the generated subgroup of G by S by $\langle S \rangle$, and the set $\{s^{-1} \mid s \in S\}$ by S^{-1}. It is clear that for every $x \in S^{-1}$, $\langle S^{-1}S \rangle = \langle xS \rangle$. If D is a subgroup of G, we write $D \leq G$. The unit element in G is denoted by e.

2. $\mathcal{L}_{\mathrm{R}}$-relation in $GM_n(R)$. We now recall some concepts for a semigroup S (see [3]). For a general semigroup S, there are five Green's relations $\mathcal{L}, \mathcal{R}, \mathcal{D}, \mathcal{H}$, and $\mathcal{T}$ for S. For instance, the definition of the relation $\mathcal{L}$ is that $a\mathcal{L}b$ if and only if $Sa \cup \{a\} = Sb \cup \{b\}$. The $\mathcal{R}$-relation is defined dually. These relations are all equivalent relations. Their equivalent classes are all called Green's classes. It can be easily proved that, for commutative semigroup, we have $\mathcal{L} = \mathcal{R} = \mathcal{D} = \mathcal{H} = \mathcal{T}$. Since $GM_n(R)$ is commutative, all the Green's relations in $GM_n(R)$ coincide with each other. So it is sufficient to discuss the Green's relation $\mathcal{L}$ in this paper. In this paper, $\mathcal{L}$ denotes the $\mathcal{L}$-relation in GM_n, and $\mathcal{L}_{\mathrm{R}}$ denotes the $\mathcal{L}$-relation in $GM_n(R)$. In this section, we will discuss the Green's relation $\mathcal{L}_{\mathrm{R}}$ in $GM_n(R)$. A necessary and sufficient condition on $\mathcal{L}_{\mathrm{R}}$ is given. For the special case of $G = \mathbf{Z}_n$ (the residue classes additive group module n), it is exactly the result of Wenchao Huang in [1]. First, some properties on $GM_n(R)$ (or GM_n) are discussed as preparation for the main investigation. For convenience, we recall some results on semigroup theory in the following (see [3]).

Let S be a semigroup, and a, $b \in S$. Then

(1) if $a \in Sb$ and $b \in Sa$, then $a\mathcal{L}b$;

(2) if $a \neq b$, then $a\mathcal{L}b$ if and only if $a \in Sb$ and $b \in Sa$.

First, we give a proposition on Abelian group.

PROPOSITION 2.1. *Let G be an n-order Abelian group. Let M be a nonempty subset of G, and $m_1, m_2, \ldots, m_l$ be l different elements in G. Then*

$$m_f M = M, \qquad f = 1, 2, \ldots, l, \tag{2.1}$$

if and only if, for $D = \langle m_1, m_2, \ldots, m_l \rangle$, there are $n_1, n_2, \ldots, n_h \in M$ such that

$$M = \dot{\bigcup}_{i=1}^{h} n_i D, \tag{2.2}$$

where $\dot{\cup}$ denotes the union of disjoint sets, that is, $n_i D \cap n_j D = \emptyset$ whenever $i \neq j, i, j = 1, 2, \ldots, h$.

Proof. "Only if": Since G is an Abelian group and D is a generated subgroup by $m_1, m_2, \ldots, m_l$ in G,

$$D = \{m_1^{t_1} m_2^{t_2} \cdots m_l^{t_l} \mid t_1, t_2, \ldots, t_l \in \mathbf{Z}\}, \tag{2.3}$$

where $\mathbf{Z}$ denotes the integer ring. We can choose $t_j \geq 0$ for $1 \leq j \leq l$. Since $M \neq \emptyset$, we can choose $n_1 \in M$. Then $n_1 D \subseteq M$. In fact, for $x \in n_1 D$, $x = n_1 m_1^{t_1} m_2^{t_2} \cdots m_l^{t_l}$, $t_j \geq 0$, $j = 1, 2, \ldots, l$. If $t_1 = 0$, then $n_1 m_1^{t_1} \in M$. If $t_1 > 0$, by (2.1), $n_1 m_1 \in M$. Then $n_1 m_1^2 = m_1(m_1 n_1) \in M, \ldots, n_1 m_1^t = m_1(n_1 m_1^{t-1}) \in M$ by (2.1). By induction l, hypothesize that $n_1 m_1^{t_1} \cdots m_{l-1}^{t_{l-l}} \in M$; then $n_1 m_1^{t_1} \cdots m_{l-1}^{t_{l-1}} m_l^{t_l} \in M$ similarly. Assume that $n_1, n_2, \ldots, n_{k-1} (k \geq 2)$ have been given such that $\dot{\cup}_{i=1}^{k-1} n_i D \subseteq M$. If $M \backslash \dot{\cup}_{i=1}^{k-1} n_i D = \emptyset$, then $M = \dot{\cup}_{i=1}^{k-1} n_i D$. Equation (2.2) holds. If $M \backslash \dot{\cup}_{i=1}^{k-1} n_i D \neq \emptyset$, we can choose $n_k \in M \backslash \dot{\cup}_{i=1}^{k-1} n_1 D$. Similarly, $n_k D \subseteq M$. By the choice of n_k, we have $n_k D \cap n_i D = \emptyset$, $i = 1, 2, \ldots, k-1$. In fact, if $n_k D \cap n_i D \neq \emptyset$, then $n_k D = n_i D$. Hence $n_k \in n_i D$. This is impossible. Since M is a finite set, by the same process, after a finite number of steps, we have $M = \dot{\cup}_{i=1}^{h} n_i D$.

"If": By (2.2) and $D = \langle m_1, m_2, \ldots, m_l \rangle$, for any f, we have $m_f n_i D = n_i m_f D = n_i D$, $i = 1, 2, \ldots, h$. Hence, $m_f M = \dot{\cup}_{i=1}^{h} m_f n_i D = \dot{\cup}_{i=1}^{h} n_i D = M$, $f = 1, 2, \ldots, l$. □

Next, we give some properties on $GM_n(R)$ or GM_n.

LEMMA 2.2. *Let $Z \neq A$, $Q \in GM_n$. Then the following are equivalent:*

(1) $QA = A$;

(2) $SN(Q)SN(A) = SN(A)$;

(3) *for every* $g \in SN(Q), gSN(A) = SN(A)$;

(4) *for every* $g \in SN(Q), P_g A = A$.

Proof. (1) $\Rightarrow$ (2): Since $P_g = P_h$ if and only if $g = h$, by the unique representation (1.5) of any nonzero element in GM_n and

$$\sum_{g \in SN(A)} P_g = A = QA = \left(\sum_{h \in SN(Q)} P_h \right) \left(\sum_{g \in SN(A)} P_g \right)$$

$$= \sum_{h \in SN(Q),\ g \in SN(A)} P_{hg} = \sum_{g \in SN(Q)SN(A)} P_g,$$

we have $SN(Q)SN(A) = SN(A)$. So (2) holds.

(2) $\Rightarrow$ (3): For every $g \in SN(Q)$, obviously $gSN(A) \subseteq SN(A)$. Since for $g_1, g_2 \in SN(A)$, $gg_1 = gg_2$ if and only if $g_1 = g_2$, it must follow $gSN(A) = SN(A)$.

(3) $\Rightarrow$ (4): For $g \in SN(Q)$, since $SN(P_g A) = gSN(A) = SN(A)$, we have $P_g A = A$.

(4) $\Rightarrow$ (1): Since $P_g A = A$ for every $g \in SN(Q)$, we have

$$A = \sum_{g \in SN(Q)} P_g A = \left(\sum_{g \in SN(Q)} P_g \right) A = QA. \qquad \square$$

PROPOSITION 2.3. *There exists an identity element in $GM_n(R)$ if and only if $|SN(R)| = 1$.*

Proof. "Only if": Assume $|SN(R)| \geq 2$, and $I = \sum_{g \in SN(I)} P_g$ is an identity element of $GM_n(R)$. Then $IR = IRE = E$. By Lemma 2.2, for $g \in SN(I)$, we have $gr_i = e, i = 1, 2, \ldots, l$, where e is a unit element in G. This implies $r_1 = r_2 = \cdots = r_l$, but that is impossible.

"If": If $|SN(R)| = 1$, then $R = P_r$ for some $r \in G$. Clearly, $I = P_r^{-1}$ is an identity element of $GM_n(R)$. □

The following result is exactly the corresponding result of Kim and Schwarz in [4] for the special case of $R = E$.

THEOREM 2.4. *If A, $B \in GM_n$, then $A\mathcal{L}B$ if and only if there exists an element $g \in G$ such that $B = P_gA$.*

Proof. "Only if": If $A\mathcal{L}B$, there exist U, $V \in GM_n$ such that $A = UB$ and $B = VA$. Then $B = VUB$. By Lemma 2.2, for every $g \in SN(VU), P_gB = B$. For $g \in SN(V)$, since $gSN(U) \subseteq SN(VU)$, so

$$B = \sum_{h \in SN(U)} P_{gh}B = P_g \left(\sum_{h \in SN(U)} P_h \right) B = P_gUB = P_gA.$$

"If": If $B = P_gA$ for some $g \in G$, then $B \in GM_nA$. On the other hand, $A = P_g^{-1}B \in GM_nB$. Therefore, $A\mathcal{L}B$. □.

For any $Z \neq R \in GM_n$, it is obvious that $A\mathcal{L}_RB$ implies $A\mathcal{L}B$. Hence, it is necessary for $A\mathcal{L}B$ that there exists P_g such that $B = P_gA$. For this reason, the following question is considered in this section: Let $Z \neq A \in GM_n(R)$; what is the necessary and sufficient condition for $P_gA\mathcal{L}_RA$ for all $g \in G$? Clearly, if $A = Z$ or H, $P_gA\mathcal{L}_RA$ holds for all $g \in G$. Hence, the above question will be discussed for the case of $A \neq Z$, H.

LEMMA 2.5. *Let $Z \neq A \in GM_n(R)$. Then $P_gA = A$ for all $g \in G$ if and only if $A = H$.*

Proof. "Only if": If $P_gA = A$ for all $g \in G$, then

$$A = \sum_{g \in G} P_gA = \left(\sum_{g \in G} P_g \right) A = HA = H.$$

"If": This is obvious. □

LEMMA 2.6. *Let $Z, H \neq A \in GM_n(R)$. Then the following are equivalent:*

(1) *$A \in GM_n(R) * A$;*

(2) *for all $g \in G, P_gA\mathcal{L}_RA$;*

(3) *there exists $g \in G$ such that $A \neq P_gA$ and $P_gA\mathcal{L}_RA$;*

(4) *there exists $g \in G$ such that $A = P_gRA$.*

Proof. (1) ⇒ (2): If $A \in GM_n(R) * A$, then there exists $B \in GM_n(R)$ such that $A = BRA$. For any $g \in G$, since $P_gA = P_gBRA = (P_gB) * A$, hence $P_gA \in MG_n(R) * A$. On the other hand, $A = (P_g^{-1}B)R(P_gA) = (P_g^{-1}B) * P_gA$, and this implies $A \in GM_n(R) * P_gA$. Hence $P_gA\mathcal{L}_RA$ for all $g \in G$.

(2) ⇒ (3): By the hypothesis $A \neq Z$, H, and Lemma 2.5, there exists $g \in G$ such that $A \neq P_gA$. By (2), then (3) holds.

(3) ⇒ (4): For some $g \in G$, $P_gA\mathcal{L}_RA$. Then $P_gA \in GM_n(R) * A$. That is, we have $P_gA = BRA$ for some $B \in GM_n(R)$. Set $C = P_g^{-1}B$, then $A = CRA$. By Lemma 2.2, for every $f \in SN(CR)$, $P_fA = A$. Since $SN(CR) = SN(C)SN(R)$, for every $g \in SN(C)$, $gSN(R) \subseteq SN(CR)$. Then, for every $g \in SN(C)$,

$$A = \sum_{r \in SN(R)} P_{gr}A = P_g \left(\sum_{r \in SN(R)} P_r \right) A = P_gRA.$$

(4) ⇒ (1): It is clear. □

THEOREM 2.7. *Let $Z, H \neq A \in GM_n(R)$. Then the following conditions are equivalent:*

(1) *for all $g \in G, P_gA\mathcal{L}_RA$;*

(2) *there exists an element* g *in* G *such that, for* $D = \langle gSN(R)\rangle$,

$$SN(A) = \dot{\cup}_{i=1}^{h} n_i D, \tag{2.4}$$

where $n_i \in SN(A), i = 1, 2, \ldots, h$.

Proof. (1) $\Rightarrow$ (2): By Lemma 2.6, there exists an element g in G such that $A = P_g RA$. Then, by Lemma 2.2, $r_i \in SN(R), 1 \le i \le l$, we have

$$gr_i SN(A) = SN(A), \qquad i = 1, 2, \ldots, l. \tag{2.5}$$

By Proposition 2.1, (2) holds.

(2) $\Rightarrow$ (1): By (2) and Proposition 2.1, (2.5) holds. Then, $SN(P_g R)SN(A) = SN(A)$ by Lemma 2.2. Therefore, $A = P_g RA$. By Lemma 2.6, (1) holds. □

Remarks. (i) According to the proof of (1) $\Rightarrow$ (2) in Lemma 2.6, (2) $\Rightarrow$ (1) in Theorem 2.7 still holds without the condition $A \neq H$.

(ii) When $A = H, SN(A) = G$. For any $g \in G$, $D = \langle gSN(R)\rangle$ is a subgroup of G. Hence, (2.4) holds.

(i) and (ii) show that Theorem 2.7 is still true without the condition $A \neq H$.

According to Lemma 2.5, Lemma 2.6, and Theorem 2.7, the following theorem on $\mathcal{L}_R$-classes is true.

THEOREM 2.8. *Let* $A \in GM_n(R)$. $\mathcal{L}_R(A)$ *denotes the* $\mathcal{L}_R$*-class of* $GM_n(R)$ *containing* A*, that is,*

$$\mathcal{L}_R(A) = \{X \mid X\mathcal{L}_R A, \quad X \in GM_n(R)\}. \tag{2.6}$$

Then either $\mathcal{L}_R(A) = \{A\}$ *or* $\{\mathcal{L}_R(A)\} \ge 2$ *with*

$$\mathcal{L}_R(A) = \{P_g A \mid g \in G\}. \tag{2.7}$$

In the following, several special cases for Theorem 2.7 are discussed.

Case 1. Suppose $|SN(R)| = 1$, and $SN(R) = \{r\}$. For arbitrary $A \in GM_n(R)$, let $SN(A) = \{n_1, n_2, \ldots, n_t\}$. Choose $g = r^{-1}$. For $D = \langle gr\rangle = \{e\}$. $SN(A) = \{n_1\} \cup \{n_2\} \cup \cdots \cup \{n_t\}$ is a union satisfying (2) of Theorem 2.7. This shows that $\mathcal{L}_R(A) = \{P_g A \mid g \in G\}$, for all $A \in GM(R)$, and leads to the following corollaries.

COROLLARY 2.9. *Let* $R = P_r, r \in G$. *If* $A, B \in GM_n(R)$, *then* $A\mathcal{L}_R B$ *if and only if* $B = P_g A$ *for some* $g \in G$. □

When $r = e$, Corollary 2.9 coincides with Theorem 2.4.

COROLLARY 2.10. *For* $A \in GM_n(R)$, *where* $R = P_r$, $r \in G$. *Then the following hold:*

(1) *if* $A \neq H, Z, |\mathcal{L}_R(A)| \ge 2$;

(2) *if* $(|SN(A)|, n) = 1, |\mathcal{L}_R(A)| = n$.

Proof. (1) Clearly.

(2) If $|\mathcal{L}_R(A)| < n$, then there exist g and h in G such that $P_g A = P_h A$ and $g \neq h$. That is, $P_{gh}^{-1} A = A$ and $gh^{-1} \neq e$. By Proposition 2.1, for $D = \langle gh^{-1}\rangle$, there are $n_1, n_2, \ldots, n_h$ such that $SN(A) = \dot{\cup}_{i=1}^{h} n_i D$. Then $|SN(A)| = h|D|$. Since $|D| > 1$ and $|D| \mid |G|$, $(|SN(A)|, n) \neq 1$. Therefore, (2) holds. □

Case 2. Suppose n is a prime and $|SN(R)| \ge 2$. Then G is a cyclic group. For any $g \in G$, since $|SN(R)| \ge 2$, we have $D = \langle gSN(R)\rangle = G$. This shows that for any Z, $H \neq A \in GM_n(R)$. A cannot satisfy condition (2) of Theorem 2.7. By Lemma 2.6, all the $\mathcal{L}_R$-classes in $GM_n(R)$ are trivial, i.e., $\mathcal{L}_R(A) = \{A\}$ for all $A \in GM_n(R)$.

Case 3. Suppose $|SN(R)| \ge 2$ and $\langle SN(R)SN(R)^{-1}\rangle = G$. Then, for any $g \in G$, $D = \langle gSN(R)\rangle = G$. By the same reasoning as the discussion in Case 2, all the $\mathcal{L}_R$-classes in $GM_n(R)$ are trivial.

3. Idempotent elements in $GM_n(R)$. Let S be a semigroup and $a \in S$. Then a is called an idempotent element of S if $a^2 = a$. In this section, we give the structure theorems for the nonzero idempotent elements of $GM_n(R)$.

LEMMA 3.1. *Let k be a positive integer, and $B_1, B_2, \ldots, B_k, \mathrm{A}$ be $k+1$ nonzero elements in $GM_n(R)$. Then the following are equivalent:*

(1) $B_1 B_2 \cdots B_k A = A$;

(2) $(\prod_{i=1}^{k} SN(B_i))SN(A) = SN(A)$;

(3) *for any $b_i \in SN(B_i), i = 1, 2, \ldots, k$, one has $SN(A) = (\prod_{i=1}^{k} b_i)SN(A)$.*

Proof. (1) $\Rightarrow$ (2): When $k = 1$, this is true by Lemma 2.2. Since $\prod_{i=1}^{k} SN(B_i) = SN(B_1 B_2 \cdots B_k)$, for any positive integer k, statement (2) holds by induction k.

(2) $\Rightarrow$ (3): For any $b_i \in SN(B_i)$, $i = 1, 2, \ldots, k$, by (2), we have $(\prod_{i=1}^{k} b_i)SN(A) \subseteq SN(A)$. Since $|(\prod_{i=1}^{k} b_i)SN(A)| = |SN(A)|, (\prod_{i=1}^{k} b_i)SN(A) = SN(A)$.

(3) $\Rightarrow$ (1): For any $b_i \in SN(B_i), i = 1, 2, \ldots, k$, write $r = \prod_{i=1}^{k} b_i$. Since $SN(P_r A) = (\prod_{i-1}^{k} b_i)SN(A) = SN(A)$ we have $P_r A = A$. Hence, we have $A = \sum_{b_1 \in SN(B_1)} \cdots \sum_{b_k \in SN(B_k)} (P_r A) = \sum_{b_1 \in SN(B_1)} \cdots \sum_{b_k \in SN(B_k)} (P_{b_1} \cdots P_{b_k} A) = (\sum_{b_1 \in SN(B_1)} P_{b_1}) \cdots (\sum_{b_k \in SN(B_k)} P_{b_k})A = B_1 \cdots B_k A$. □

LEMMA 3.2. *Let $Z \neq I \in GM_n(R)$. Then the following are equivalent:*

(1) *I is an idempotent element in $GM_n(R)$;*

(2) *for $D = \langle SN(I)SN(R) \rangle$, one has*

$$SN(I) = \dot{\cup}_{i=1}^{h} n_i D, \qquad n_i \in SN(I), \quad i = 1, 2, \ldots, h. \tag{3.1}$$

Proof. I being an idempotent element in $GM_n(R)$ means $IRI = I$. By Lemma 3.1, $IRI = I$ if and only if for any $n_i \in SN(I), r_j \in SN(R)$

$$n_i r_j SN(I) = SN(I). \tag{3.2}$$

By Proposition 2.1, (3.2) holds if and only if (2) holds. □

THEOREM 3.3. *Let $Z \neq I \in GM_n(R)$. Then the following are equivalent:*

(1) *I is an idempotent element in $GM_n(R)$;*

(2) *for $D = \langle SN(I)SN(R) \rangle$, one has*

$$SN(I) = n_i D, \quad \textit{for some } n_i \in SN(I). \tag{3.3}$$

Proof. (1) $\Rightarrow$ (2): Suppose I is an idempotent element in $GM_n(R)$. By Lemma 3.2, for $D = \langle SN(I)SN(R) \rangle$, we have $SN(I) = \dot{\cup}_{i=1}^{h} n_i D$, where $n_i \in SN(I), i = 1, 2, \ldots, h$. We will prove $h = 1$. Suppose $h \geq 2$. Since $n_1 r_1, n_2 r_1 \in D, n_1 n_2^{-1} \in D$. Then $n_1 D = n_2 D$, which contradicts $n_1 D \cap n_2 D = \emptyset$. Hence, $h = 1$ and (2) holds.

(2) $\Rightarrow$ (1): This is the conclusion of Lemma 3.2. □

THEOREM 3.4 (the first structure theorem for idempotent elements). *Suppose $e \in SN(R)$. Let $\langle SN(R) \rangle \leq D \leq G$. Then*

$$I = \sum_{g \in D} P_g \tag{3.4}$$

is a nonzero idempotent element in $GM_n(R)$. All the nonzero idempotent elements in $GM_n(R)$ are obtained in this manner.

Proof. (a) Since $\langle SN(R) \rangle \leq D$ and $SN(I) = D$, $\langle SN(R)SN(I) \rangle = \langle (SN(I) \rangle = D$. By Theorem 3.3, I is an idempotent element in $GM(R)$.

(b) Conversely, let I be an arbitrary nonzero idempotent element in $GM_n(R)$. By Theorem 3.3, for some $n_i \in SN(I)$, $SN(I) = n_i D$, where $D = \langle SN(I)SN(R) \rangle$. Since

$e \in SN(R)$, $n_i \in D$. Hence, $SN(I) = n_i D = D$ and $\langle SN(R)\rangle \le D \le G$. The proof is completed. □

Let $D \le G$. D^+ denotes the cardinality of the set $\{F \mid D \le F \le G\}$. Then, by Theorem 3.4, we have the following result.

COROLLARY 3.5. *If $e \in SN(R)$ and $D = \langle SN(R)\rangle$, there exist exactly D^+ different nonzero idempotent elements in $GM_n(R)$.* □

In the following, we discuss the idempotent elements in $GM_n(R)$ for the general case (without the condition $e \in SN(R)$).

LEMMA 3.6. *Let $I \in GM_n(R)$. If $SN(I) = D$, where $\langle SN(R)\rangle \le D \le G$, then I is an idempotent element in $GM_n(R)$.*

Proof. The proof is the same as part (a) of the proof of Theorem 3.4. □

THEOREM 3.7. *Let $D = \langle SN(R)\rangle$. Then there exist exactly D^+ different nonzero idempotent elements in $GM_n(R)$ satisfying $e \in SN(I)$.*

Proof. By Lemma 3.6, there exist at least D^+ different nonzero idempotent elements in $GM_n(R)$ satisfying $e \in SN(I)$. Conversely, let I be a nonzero idempotent element in $GM_n(R)$ satisfying $e \in SN(I)$. By Theorem 3.3, for some $n_i \in SN(I)$. $SN(I) = n_i D'$, where $D' = \langle SN(I)SN(R)\rangle$. Then $e = n_i d \in D'$, $d \in D'$, and $n_i^{-1} \in D'$. So $n_i \in D'$. Hence, $SN(I) = D'$. Since $e \in SN(I)$, $D = \langle SN(R)\rangle \le D'$. This shows that each nonzero idempotent element I in $GM_n(R)$ satisfying $e \in SN(I)$ corresponds to a subgroup D' of G satisfying $D \le D'$. Combining the above discussion, we have proved this theorem. □

THEOREM 3.8. *Let $Z \ne I \in GM_n(R)$ and $e \notin SN(I)$. Then the following are equivalent:*

(1) *I is an idempotent element in $GM_n(R)$;*

(2) *there exists a nonunit element g in G such that, for a subgroup D of G satisfying $\langle gSN(R)\rangle \le D$,*

$$SN(I) = gD. \tag{3.5}$$

Proof. (1) ⇒ (2): Suppose I is an idempotent element in $GM_n(R)$. By Theorem 3.3, for $D = \langle SN(I)SN(R)\rangle$, there exists g in G such that $SN(I) = gD$. Since $e \notin SN(I)$, we have $e \ne g \in SN(I)$. Obviously $\langle gSN(R)\rangle \le D$.

(2) ⇒ (1): Let $D' = \langle gDSN(R)\rangle$. Because $\langle gSN(R)\rangle \le D$, $D' = \langle gSN(R)D\rangle = \langle D\rangle = D$. □

THEOREM 3.9 (the second structure theorem for idempotent elements). *Let $V = \langle SN(R)^{-1}SN(R)\rangle$. Let D be a subgroup of G satisfying $V \le D$. Then*

$$I = \sum_{g \in cD} P_g, \quad \text{for some } c \in SN(R)^{-1}, \tag{3.6}$$

is a nonzero idempotent element in $GM_n(R)$. All the nonzero idempotent elements in $GM_n(R)$ are obtained in this manner.

Proof. (a) If $c = e$, then by Lemma 3.6, $I = \sum_{g \in D} P_g$ is an idempotent element in $GM_n(R)$. If $c \ne e$, then $e \notin SN(I)$ and $V = \langle cSN(R)\rangle$. Hence, by Theorem 3.8, $I = \sum_{g \in cD} P_g$ is an idempotent element in $GM_n(R)$.

(b) Conversely, let I be a nonzero idempotent element in $GM_n(R)$. If $e \in SN(I)$, then, by the proof of Theorem 3.7, $SN(I) = D$, where D is a subgroup of G satisfying $\langle SN(R)\rangle \le D$. If $e \notin SN(I)$, by Theorem 3.8, there exists $e \ne g \in G$ such that, for a subgroup D of G satisfying $\langle gSN(R)\rangle \le D$, $SN(I) = gD$. Since $V = \langle SN(R)^{-1}SN(R)\rangle \le \langle hSN(R)\rangle \le D$, for $h = e$ or g, $hr \in D$ for any $r \in SN(R)$.

Then $r^{-1}D = hD$. That is, there exists a subgroup D of G satisfying $V \leq D$ such that $I = \sum_{g \in cD} P_g$ for some $c \in SN(R)^{-1}$. □

The idempotent element obtained in Theorem 3.9 is denoted $I(c, D)$. Obviously, $I(c, G) = H$ for any possible c.

LEMMA 3.10. *In $GM_n(R), I(c, D) = I(c', D')$ if and only if $D = D'$.*

Proof. "Only if": Suppose $I(c, D) = I(c', D')$. Since $SN(I(c, D)) = SN(I(c', D'))$, we have $cD = c'D'$. Hence $c^{-1}c'D' = D$. Since $e \in D, e = c^{-1}c'd$ for some $d \in D'$. Then $c^{-1}c' = d^{-1} \in D'$. So $c^{-1}c'D' = D'$. Therefore, $D = D'$.

"If": If $D = D'$, since c, $c' \in SN(R)^{-1}$, $c^{-1}c' \in V$, and $V \leq D$, so $c^{-1}c' \in D$. Then $cD = c'D = c'D'$. Therefore, $I(c, D) = I(c', D')$. □

The set of all the idempotent elements in $GM_n(R)$ is denoted by $Id(GM_n(R))$. We have the following counting theorem on $Id(GM_n(R))$.

THEOREM 3.11. *Let $Z \neq R \in GM_n(R)$. If $V = \langle SN(R)^{-1}SN(R) \rangle$, then*

$$Id(GM_n(R)) = V^+ + 1. \tag{3.7}$$

Proof. By Theorem 3.9 and Lemma 3.10, the number of nonzero idempotent elements in $GM_n(R)$ is exactly V^+. Clearly, Z is also an idempotent element in $GM_n(R)$. □

4. Maximal subgroups in $GM_n(R)$. Let I be an idempotent element in $GM_n(R)$. The maximal subgroup of $GM_n(R)$ containing I is denoted by $G_I(R)$ (the definition of the maximal subgroup of a semigroup can be seen in [3]). In this section, we will investigate the structure of $G_I(R)$. Since $GM_n(R)$ is commutative, all the Green's relations for $GM_n(R)$ coincide with each other. Therefore, for any idempotent element I in $GM_n(R)$, we have $G_I(R) = \mathcal{L}_R(I)$.

LEMMA 4.1. *If I is an idempotent element in $GM_n(R)$, then for any $g \in G$, one has $P_gI \in \mathcal{L}_R(I)$.*

Proof. If $I = Z$, the conclusion is obviously true. Suppose $I \neq Z$. It is sufficient to prove

$$P_gI \in GM_n(R) * I \text{ and } I \in GM_n(R) * (P_gI), \text{ for any } g \in G. \tag{4.1}$$

Since $I = IRI$, we have $P_gI = (P_gI)RI$ and $I = (P_g^{-1}I)R(P_gI)$ for any $g \in G$. This shows that (4.1) is true. □

For A, $B \in GM_n(R)$, clearly, if A and B are $\mathcal{L}$-equivalent in $GM_n(R)$, then A and B are $\mathcal{L}$-equivalent in GM_n. According to Theorem 2.4 and Lemma 4.1, for any idempotent element I in $GM_n(R)$, we have

$$\mathcal{L}_R(I) = \{P_gI \mid g \in G\}. \tag{4.2}$$

THEOREM 4.2 (structure theorem for maximal subgroups). *Let $I = I(c, D) = \sum_{g \in cD} P_g$ be a nonzero idempotent element in $GM_n(R)$ obtained in the manner of Theorem 3.9, and*

$$G = \dot{\cup}_{i=1}^{s} g_i D. \tag{4.3}$$

Then

$$G_I(R) = \{P_{g_i}I \mid i = 1, 2, \ldots, s\} \tag{4.4}$$

and

$$|G_I(R)| = |G|/|D|. \tag{4.5}$$

Proof. By the above investigation, $G_I(R) = \{P_gI \mid g \in G\}$. To prove (4.4) and (4.5), it is sufficient to prove the following:

(1) $gD = hD$ if and only if $P_gI = P_hI$;

(2) for every i, if $g, h \in g_iD$, $P_gI = P_hI$.

"Proof of (1)": First, we have $P_gI = P_g \sum_{k\in cD} P_k = \sum_{k\in cD} P_{gk} = \sum_{k\in gcD} P_k$. Hence, for $g, h \in G$,

$$P_gI = P_hI \Longleftrightarrow \sum_{k\in gcD} P_k = \sum_{k\in hcD} P_k \Longleftrightarrow gcD = hcD \Longleftrightarrow gD = hD.$$

"Proof of (2)": For every i, if $g, h \in g_iD$, then $gD = g_iD = hD$. By (1), we have $P_gI = P_hI$. □

THEOREM 4.3. *Let I satisfy the conditions of Theorem 4.2. Then $G_I(R) \cong G/D$ (group isomorphism), where G/D denotes the quotient group of G module D.*

Proof. Set

$$\Phi : G_I(R) \to G/D; \qquad P_gI \to gD.$$

By the proof of Theorem 4.2, it is clear that Φ is a bijection from $G_I(R)$ to G/D. And, since $IRI = I$, for $g, h \in G$, $\Phi(P_gI * P_hI) = \Phi(P_gIRP_hI) = \Phi(P_{gh}I) = ghD = (gD)(hD) = \Phi(P_gI)\Phi(P_hI)$. Hence, Φ is an isomorphism from $G_I(R)$ to G/D. □

The following theorem gives an answer for the problem about trivial maximal subgroups in $GM_n(R)$.

THEOREM 4.4. *For arbitrary $Z \neq R \in GM_n$, $GM_n(R)$ contains exactly two trivial maximal subgroups, which are $G_Z(R)$ and $G_H(R)$.*

Proof. Let $I = I(c, D)$ be a nonzero idempotent element determined in the manner of Theorem 3.9. If $G_I(R)$ is trivial, then, by Theorem 4.2, $|G_I(R)| = |G|/|D| = 1$. So $G = D$ and $cD = D = G$. Hence, $I = \sum_{g\in G} P_g = H$. Therefore, $G_Z(R)$ and $G_H(R)$ are the only trivial maximal subgroups in $GM_n(R)$. □

Remark. If $Z \neq R \in GM_n$ and $|SN(R)| \geq 2$, let $I = I(c, D)$ be a nonzero idempotent element in $GM_n(R)$ obtained in the manner of Theorem 3.9; then $|G_I(R)| < n$. In fact, suppose $|G_I(R)| = n$, by Theorem 4.2, $D = \{e\}$. Since $\langle cSN(R)\rangle \leq D$, $\langle cSN(R)\rangle = \{e\}$. Then $SN(R) = \{c^{-1}\}$, and $I = P_c$, which contradicts $|SN(R)| \geq 2$.

5. Regular elements in $GM_n(R)$. In this section, we will determine all the regular elements in $GM_n(R)$. Let S be a semigroup and $a \in S$. Then a is said to be a regular element of S if $axa = a$ for some $x \in S$. By referring to [3], we can obtain the following:

(1) if $A \in GM_n(R)$ is regular, then all the elements in $\mathcal{L}_R(A)$ are regular;

(2) $A \in GM_n(R)$ is regular if and only if $\mathcal{L}_R(A)$ contains an idempotent element.

According to (1) and (2), and the results in sections 3 and 4, we can obtain the following theorem immediately.

THEOREM 5.1. *Let $A \in GM_n(R)$. Then the following are equivalent:*

(1) *A is a regular element in $GM_n(R)$;*

(2) *there exist g in G and I in $GM_n(R)$, where I is an idempotent element of $GM_n(R)$, such that $A = P_gI$;*

(3) *there exists an idempotent element I of $GM_n(R)$ such that $A \in G_I(R)$.* □

The set of all the regular elements in $GM_n(R)$ is denoted by $Reg(GM_n(R))$.

THEOREM 5.2. *For $GM_n(R)$, the following are true:*

(1) $Reg(GM_n(R)) = \dot{\cup}_{I\in Id(GM_n(R))} G_I(R)$;

(2) $Reg(GM_n(R)) = 1 + \sum_{V\leq D\leq G} |G|/|D|$, *where* $V = \langle SN(R)^{-1}SN(R)\rangle$.

Proof. (1) By Theorem 5.1, we have $Reg(GM_n(R)) = \cup_{I\in Id(GM_n(R))}G_I(R)$. If $G_{I_1}(R)\cap G_{I_2}(R)\neq\emptyset$, there are g and h in G such that $P_gI_1 = P_hI_2$. Then $I_1 = P_{g^{-1}h}I_2 \in G_{I_2}(R)$, $I_2 = P_{gh^{-1}}I_1 \in G_{I_1}(R)$. Hence, $G_{I_1}(R)\subseteq G_{I_2}(R)$ and $G_{I_2}(R)\subseteq G_{I_1}(R)$. Therefore, $G_{I_1}(R) = G_{I_2}(R)$. Thus, we have $G_{I_1}(R)\cap G_{I_2}(R) = \emptyset$ if and only if $I_1 = I_2$. Hence, $Reg(GM_n(R)) = \dot\cup_{I\in Id(GM_n(R))}G_I(R)$.

(2) By (1) and Theorem 4.2, we have that $|Reg(GM_n(R))| = \sum_{I\in Id(GM_n(R))}|G_I(R)| = 1+\sum_{I\in Id(GM_n(R)),I\neq Z}|G_I(R)| = 1+\sum_{V\le D\le G}|G|/|D|$. That is, (2) holds.

6. Algorithm and examples.

6.1. $\mathcal{L}_R$-classes.

Example 1. Let $n = 18, G = \mathbf{Z}_1+\mathbf{Z}_2+\mathbf{Z}_3$, and $R = \sum_{g\in SN(R)}P_g$, where

$$SN(R) = \{(0,1,0),(0,1,1),(0,1,2),(1,1,1)\ (1,1,2)\}.$$

Suppose $A = \sum_{g\in SN(A)}P_g$, where $SN(A) = \{(x,y,z)\mid x = 0,1; y = 0,1; z = 0,1,2\}$. Choose $g = (0,2,0)$. For $D = \langle gSN(R)\rangle = \langle(0,0,0),(0,0,1),(0,0,2),(1,0,1),(1,0,2)\rangle = \{(x,0,z)\mid x = 0,1; z = 0,1,2\}$. We have

$$SN(A) = D\ \dot\cup\ (0,1,0)D.$$

By Theorem 2.7, $\mathcal{L}_R(A) = \{P_gA\mid g\in G\}$. By calculating,

$$\mathcal{L}_R(A) = \{A, P_{(0,1,0)}A,\quad P_{(0,2,0)}A\}.$$

Example 2. Let Z, $H\neq A\in GM_n(R)$. If for every $g\in G|\langle gSN(R)\rangle|$ is not a divisor of $|SN(A)|$, then $\mathcal{L}_R(A) = \{A\}$.

6.2. Idempotent elements. According to Theorem 3.9, Lemma 3.10, and Theorem 3.11, we can find all the idempotent elements in $GM_n(R)$ for any $Z\neq R\in GM_n$. An algorithm for obtaining all the idempotent elements in $GM_n(R)$ is given as follows.

Algorithm. Let Z, $H\neq R\in GM_n(R)$.

Step 1. Compute $V = \langle SN(R)^{-1}SN(R)\rangle$.

Step 2. Compute D satisfying $V\le D\le G$, say $D_1, D_2, \ldots, D_k$.

Step 3. Choose $c_i = r_{t_i}^{-1}$, where $r_{t_i}\in SN(R)$, $1\le t_i\le l$, $i = 1,2,\ldots,k$.

Step 4. Form all idempotent elements of

$$Id(GM_n(R)) = \{Z\}\cup\left\{I(c_i,D_i) = \sum_{g\in c_iD_i}P_g\mid i = 1,2,\ldots,k\right\}.$$

Example 3. Let $n = 60, G = \mathbf{Z}_2+\mathbf{Z}_2+\mathbf{Z}_{15}$, and $SN(R) = \{(1,0,7),(0,1,8)\}$. We will determine all the idempotent elements in $GM_n(R)$.

Step 1. $V = \langle(0,0,0),(1,1,14),(1,1,1)\rangle = \{(x,x,y)\mid x = 0,1; y = 0,1,\ldots,14\}$.

Step 2. All the subgroups D of G satisfying $V\le D$ are V and G.

Step 3. Choose $c_1 = (1,0,8), c_2 = (1,0,8)$.

Step 4. $c_1V = \{(1,0,x),(0,1,y)\mid x,y = 0,1,2,\ldots,14\}$ and $c_2G = G$. Hence,

$$Id(GM_n(R)) = \left\{Z, H, \sum_{g\in(1,0,8)V}P_g\right\}.$$

We obtain exactly $V^+ = 2$ nonzero idempotent elements in $GM_n(R)$.

If $e \in SN(R)$, according to Theorem 3.4, we can also find out all the nonzero idempotent elements in $GM_n(R)$. This can be demonstrated by the following example.

Example 4. Let $n = 50, G = \mathbf{Z}_5+\mathbf{Z}_5+\mathbf{Z}_2$, and $SN(R) = \{(0, y, 0) \mid y = 0, 1, 2, 3\}$. Then, $\langle SN(R)\rangle = \{(0, y, 0) \mid y = 0, 1, 2, 3, 4\}$. By calculating, all subgroups D of G satisfying $\langle SN(R)\rangle \leq D$ are D_1, D_2, D_3, and D_4 where

$$D_1 = \langle SN(R)\rangle, \quad D_2 = \{(x, y, 0) \mid x = 0, 1, 2, 3, 4; y = 0, 1, 2, 3, 4\},$$
$$D_3 = \{(0, y, z) \mid y = 0, 1, 2, 3, 4; z = 0, 1\}, \quad \text{and } D_4 = G.$$

By Theorem 3.4, there exist exactly four nonzero idempotent elements in $GM_n(R)$ as follows:

$$I_1 = P_{(0,0,0)} + P_{(0,1,0)} + P_{(0,2,0)} + P_{(0,3,0)} + P_{(0,4,0)},$$
$$I_2 = \sum_{g\in D_2} P_g, \quad I_3 = \sum_{g\in D_3} P_g, \quad I_4 = H.$$

6.3. Maximal subgroups. According to Theorem 4.2, all the maximal subgroups of $GM_n(R)$ containing a nonzero idempotent element can be determined; considering $G_Z(R)$, there are exactly $V^+ + 1$ maximal subgroups in $GM_n(R)$, where $V = \langle SN(R)^{-1}SN(R)\rangle$.

Example 5. Let $n = 200$, $G = \mathbf{Z}_2 + \mathbf{Z}_4 + \mathbf{Z}_5 + \mathbf{Z}_5$, and $SN(R) = \{(1, 0, 0, 0), (1, 1, 1, 1)\}$. Then, $V = \langle(0, 0, 0, 0), (0, 3, 4, 4), (0, 1, 1, 1)\rangle = \{(0, x, y, y) \mid x = 0, 1, 2, 3; y = 0, 1, 2, 3, 4\}$. By calculating, we obtain

$$D_1 = V, \quad D_2 = \{(0, x, y, z) \mid x = 0, 1, 2, 3; y, z = 0, 1, 2, 3, 4\},$$
$$D_3 = \{(x, y, z, z) \mid x = 0, 1; y = 0, 1, 2, 3; z = 0, 1, 2, 3, 4\}, \quad D_4 = G.$$

Choose $c = (1, 0, 0, 0)$. Then

$$Id(GM_n(R)) = \left\{ Z, H, \sum_{x=0}^{3}\sum_{y=0}^{4} P_{(1,x,y,y)}, \sum_{x=0}^{3}\sum_{y=0}^{4}\sum_{z=0}^{4} P_{(1,x,y,z)}, \right.$$
$$\left. \sum_{x=0}^{1}\sum_{y=0}^{3}\sum_{z=0}^{4} P_{(x,y,z,z)} \right\}.$$

So there are five maximal subgroups in $GM_n(R)$ as follows:

$$G_Z(R) = \{Z\}, \qquad G_H(R) = \{H\},$$
$$G_{I(c,V)}(R) = \{P_{(x,0,y,0)}I(c, V) \mid x = 0, 1; y = 0, 1, 2, 3, 4\},$$
$$G_{I(c,D_2)}(R) = \{P_{(x,0,0,0)}I(c, D_2) \mid x = 0, 1\},$$
$$G_{I(c,D_3)}(R) = \{P_{(0,0,0,z)}I(c, D_3) \mid z = 0, 1, 2, 3, 4\}.$$

Clearly, $G_Z(R) \cong G_H(R) = \{e\}$, $G_{I(c,V)}(R) \cong \mathbf{Z}_2 + \mathbf{Z}_5, G_{I(c,D_2)}(R) \cong \mathbf{Z}_2$, $G_{I(c,D_3)}(R) \cong \mathbf{Z}_5$.

6.4. Regular elements. According to Theorem 5.2, we can determine all the regular elements in $GM_n(R)$ by computing all the idempotent elements in $GM_n(R)$.

Example 6. Let $n = 200$, $G = \mathbf{Z}_2 + \mathbf{Z}_4 + \mathbf{Z}_5 + \mathbf{Z}_5$, and $SN(R) = \{(1, 0, 0, 0), (1, 1, 1, 1)\}$. Then, $V = \{(0, x, y, y) \mid x = 0, 1, 2, 3; y = 0, 1, 2, 3, 4\}$. By Theorem 5.2,

$$|Reg(GM_n(R))| = 1 + \sum_{V\leq D\leq G} |G|/|D|$$
$$= 1 + 10 + 2 + 5 + 1 = 19.$$

So there exist exactly 19 regular elements in $GM_n(R)$ which are the matrices in all the maximal subgroups $G_Z(R)$, $G_H(R)$, $G_{I(c,V)}(R)$, $G_{I(c,D_2)}(R)$, and $G_{I(c,D_3)}(R)$ in Example 5.

Acknowledgment. The author thanks the referees and Professor George P. H. Styan for their helpful suggestions to this paper.

REFERENCES

[1] W. C. Huang, *On the sandwich semigroups of circulant Boolean matrices*, Linear Algebra Appl., 179 (1993), pp. 135–160.
[2] K. Wang, *On the generalizations of circulants*, Linear Algebra Appl., 25 (1979), pp. 197–218.
[3] A. H. Clifford and G. B. Preston, *The Algebraic Theory of Semigroups*, Vol. 1, Amer. Math. Soc., Providence, RI, 1961.
[4] K. H. Kim and S. Schwarz, *The semigroup of circulant Boolean matrices*, Czechoslovak Math. J., 26 (1976), pp. 632–635.
[5] J. M. Cen, *Group matrix ring and its applications*, J. Ningbo Normal University, 17 (1995), pp. 11–17.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 429–443, April 1998

010

PERTURBATION ANALYSES FOR THE CHOLESKY DOWNDATING PROBLEM*

XIAO-WEN CHANG† AND CHRISTOPHER C. PAIGE†

Abstract. New perturbation analyses are presented for the block Cholesky downdating problem $U^TU = R^TR - X^TX$. These show how changes in R and X alter the Cholesky factor U. There are two main cases for the perturbation matrix ΔR in R: (1) ΔR is a general matrix; (2)ΔR is an upper triangular matrix. For both cases, first-order perturbation bounds for the downdated Cholesky factor U are given using two approaches — a detailed "matrix–vector equation" analysis which provides tight bounds and resulting true condition numbers, which unfortunately are costly to compute, and a simpler "matrix equation" analysis which provides results that are weaker but easier to compute or estimate. The analyses more accurately reflect the sensitivity of the problem than previous results. As $X \to 0$, the asymptotic values of the new condition numbers for case (1) have bounds that are independent of $\kappa_2(R)$ if R was found using the standard pivoting strategy in the Cholesky factorization, and the asymptotic values of the new condition numbers for case (2) are unity. Simple reasoning shows this last result must be true for the sensitivity of the problem, but previous condition numbers did not exhibit this.

Key words. perturbation analysis, sensitivity, condition, asymptotic condition, Cholesky factorization, downdating

AMS subject classifications. 15A23, 65F35

PII. S0895479896304113

1. Introduction.

Let $A \in \mathcal{R}^{n\times n}$ be a symmetric positive definite matrix. Then there exists a unique upper triangular matrix $R \in \mathcal{R}^{n\times n}$ with positive diagonal elements such that $A = R^TR$. This factorization is called the Cholesky factorization, and R is called the Cholesky factor of A (see, for example, [13]).

In this paper we give perturbation analyses of the following problem: given an upper triangular matrix $R \in \mathcal{R}^{n\times n}$ and a matrix $X \in \mathcal{R}^{k\times n}$ such that $R^TR - X^TX$ is positive definite, find an upper triangular matrix $U \in \mathcal{R}^{n\times n}$ with positive diagonal elements such that

$$U^TU = R^TR - X^TX. \tag{1.1}$$

This problem is called the block Cholesky downdating problem, and the matrix U is referred to as the downdated Cholesky factor. The block Cholesky downdating problem has many important applications, and it has been extensively studied in the literature (see [1, 2, 3, 8, 9, 11, 12, 16, 17, 18]).

Perturbation results for the single Cholesky downdating problem ($k = 1$) were presented by Stewart [18]. Eldén and Park [10] made an analysis for block downdating. But these two papers just considered the case that only R or X is perturbed. More complete analyses, with both R and X being perturbed, were given by Pan [15] and Sun [20]. Pan [15] gave first-order perturbation bounds for single downdating. Sun [20] gave strict, also first-order perturbation bounds for single downdating and first-order perturbation bounds for block downdating.

*Received by the editors May 7, 1996; accepted for publication (in revised form) by G. P. Styan April 2, 1997. This research was partially supported by NSERC of Canada grant OGP0009236.
http://www.siam.org/journals/simax/19-2/30411.html
†School of Computer Science, McGill University, Montreal, Quebec H3A 2A7, Canada (chang@cs.mcgill.ca, chris@cs.mcgill.ca).

The main purpose of this paper is to establish new first-order perturbation results and present new condition numbers which more closely reflect the true sensitivity of the problem. In section 2 we will give the key result of Sun [20] and a new result using the approach of these earlier papers. In section 3 we present new perturbation results, first by the straightforward matrix equation approach, then by the more detailed and tighter matrix–vector equation approach. The basic ideas behind these two approaches were discussed by Chang, Paige, and Stewart [6, 7]. We give numerical results and suggest practical condition estimators in section 4.

Previous papers implied the change ΔR in R was upper triangular, and Sun [20] said this, but neither he nor the others made use of this. In fact a backward stable algorithm for computing U given R and X would produce the exact result $U_c = U + \Delta U$ for nearby data $R + \Delta R$ and $X + \Delta X$, where it is not clear that ΔR would be upper triangular — the form of the equivalent backward error ΔR would depend on the algorithm, and if it were upper triangular, it would require a rounding error analysis to show this. Thus, for completeness it seems necessary to consider two separate cases—upper triangular ΔR and general ΔR. We do this throughout sections 3 and 4, and get stronger results for upper triangular ΔR than in the general case.

In any perturbation analysis it is important to examine how good the results are. In section 3.2 we produce provably tight bounds, leading to the true condition numbers (for the norms chosen). The numerical example in section 4 indicates how much better the results of this new analysis can be compared with some earlier ones, but a theoretical understanding is also desirable. By considering the asymptotic case as $X \to 0$, the results simplify, and are easily understandable. We show the new results have the correct properties as $X \to 0$, in contrast to earlier results.

Before proceeding, let us introduce some notation. Let $B = (b_{ij}) \in \mathcal{R}^{n\times n}$; then $\mathrm{up}(B)$, $\mathrm{sut}(B)$, $\mathrm{slt}(B)$, and $\mathrm{diag}(B)$ are defined by

$$\mathrm{up}(B) \equiv \begin{pmatrix} \frac{1}{2}b_{11} & b_{12} & \cdot & b_{1n} \\ 0 & \frac{1}{2}b_{22} & \cdot & b_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \frac{1}{2}b_{nn} \end{pmatrix}, \tag{1.2}$$

$$\mathrm{sut}(B) \equiv \begin{pmatrix} 0 & b_{12} & b_{13} & \cdot & b_{1n} \\ 0 & 0 & b_{23} & \cdot & b_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & 0 & b_{n-1,n} \\ 0 & \cdot & \cdot & \cdot & 0 \end{pmatrix}, \qquad \mathrm{slt}(B) \equiv \mathrm{sut}(B^T)^T, \tag{1.3}$$

and

$$\mathrm{diag}(B) = \mathrm{diag}(b_{11}, b_{22}, \ldots, b_{nn}). \tag{1.4}$$

2. Basics, previous results, and an improvement. Let Γ satisfy $\Gamma^T\Gamma = I_n - R^{-T}X^TXR^{-1}$ (so Γ could be the Cholesky factor of $I_n - R^{-T}X^TXR^{-1}$), and let $\sigma_n(\Gamma)$ be the smallest singular value of Γ. Notice that for fixed R, $\Gamma^T\Gamma \to I_n$ as $X \to 0$, so $\sigma_n(\Gamma) \to 1$. First we derive some relationships among U, R, X, and Γ.

1) From (1.1) obviously we have

$$\|U\|_2 \le \|R\|_2, \qquad \|X\|_2 \le \|R\|_2. \tag{2.1}$$

2) From (1.1) it follows that

$$RU^{-1}U^{-T}R^T = (I_n - R^{-T}X^TXR^{-1})^{-1},$$

so that taking the 2-norm gives

$$\|RU^{-1}\|_2 = \frac{1}{\sigma_n(\Gamma)}. \tag{2.2}$$

3) From (1.1) we have

$$U^{-T}X^TXU^{-1} = U^{-T}R^TRU^{-1} - I_n,$$

which, combined with (2.2), gives

$$\|XU^{-1}\|_2 = \sqrt{\|RU^{-1}\|_2^2 - 1} = \frac{\sqrt{1-\sigma_n^2(\Gamma)}}{\sigma_n(\Gamma)} = \sqrt{1-\sigma_n^2(\Gamma)}\|RU^{-1}\|_2. \tag{2.3}$$

4) From (1.1) we have

$$R^{-T}X^TXR^{-1} = I_n - R^{-T}U^TUR^{-1},$$

which, combined with (2.2), gives

$$\|XR^{-1}\|_2 = \sqrt{1-\sigma_{\min}^2(UR^{-1})} = \sqrt{1-\frac{1}{\|RU^{-1}\|_2^2}} = \sqrt{1-\sigma_n^2(\Gamma)}. \tag{2.4}$$

5) By (2.2) we have

$$\|R\|_2 = \|RU^{-1}U\|_2 \le \|RU^{-1}\|_2\|U\|_2 = \frac{\|U\|_2}{\sigma_n(\Gamma)}, \tag{2.5}$$

$$\|U^{-1}\|_2 = \|R^{-1}RU^{-1}\|_2 \le \|R^{-1}\|_2\|RU^{-1}\|_2 = \frac{\|R^{-1}\|_2}{\sigma_n(\Gamma)}. \tag{2.6}$$

6) Finally, from (2.4) we see

$$\frac{\|X\|_2}{\|R\|_2} \le \|XR^{-1}\|_2 = \sqrt{1-\sigma_n^2(\Gamma)}. \tag{2.7}$$

The condition number for the single Cholesky downdating problem ($k = 1$, $X^T = x$ say) suggested by Pan [15] is

$$\beta_0 \equiv c_n \cdot \sqrt{2}\kappa_2(R)\left(\frac{\kappa_2(R)\|v\|_2^2}{1-\|v\|_2^2} + 1\right),$$

where $\sqrt{2}n \le c_n \le \sqrt{2}n^{3/2}$, $\kappa_2(R) \equiv \|R\|_2\|R^{-1}\|_2$, and $v \equiv R^{-T}x \in \mathcal{R}^n$. The condition number for block downdating proposed by Sun [20] is

$$\beta_1 \equiv \sqrt{2}\,\frac{\kappa_2(U)}{\sigma_n^2(\Gamma)}.$$

In particular, when $k = 1$, from (2.4) we see $\sigma_n(\Gamma) = \sqrt{1-\|v\|_2^2}$, and thus

$$\beta_1 = \sqrt{2}\,\frac{\kappa_2(U)}{1-\|v\|_2^2}.$$

In this case, when $\|v\|$ is sufficiently close to zero, β_0/c_n and β_1 are approximately equal; otherwise it is difficult to assess which one is smaller. But [20] indicated that numerical tests show in most cases β_1 is smaller than β_0/c_n, and that when $\kappa_2(R) \gg 1$ and/or $1 - \|v\|_2^2 \approx 0$, the former is much smaller than the latter.

Now we use a similar approach to Sun's to derive a new bound for block downdating, from which Sun's bound follows.

To derive first-order perturbation results we consider the perturbed version of (1.1):

$$(U + \Delta U)^T(U + \Delta U) = (R + \Delta R)^T(R + \Delta R) - (X + \Delta X)^T(X + \Delta X), \tag{2.8}$$

where U, $U + \Delta U$, and R are upper triangular matrices with positive diagonal elements. Clearly when ΔR and ΔX are sufficiently small, (2.8) has a unique solution ΔU. Multiplying out the two sides of (2.8) and ignoring second-order terms, we obtain a linear matrix equation for the first-order approximation $\widehat{\Delta U}$ to ΔU:

$$U^T\widehat{\Delta U} + \widehat{\Delta U}^T U = R^T\Delta R + \Delta R^T R - X^T\Delta X - \Delta X^T X. \tag{2.9}$$

In fact it is straightforward to show $\widehat{\Delta U} = \dot{U}(0)$, the rate of change of $U(\tau)$ at $\tau = 0$, where $U(\tau)^T U(\tau) \equiv (R + \tau\Delta R)^T(R + \tau\Delta R) - (X + \tau\Delta X)^T(X + \tau\Delta X)$, $0 \le \tau \le 1$, so $\widehat{\Delta U}$ also has a precise meaning. From (2.9) we have

$$\widehat{\Delta U}U^{-1} + (\widehat{\Delta U}U^{-1})^T = U^{-T}(R^T\Delta R + \Delta R^T R - X^T\Delta X - \Delta X^T X)U^{-1}.$$

Notice since $\widehat{\Delta U}U^{-1}$ is upper triangular, it follows, with (1.2), that

$$\widehat{\Delta U} = \mathrm{up}[U^{-T}(R^T\Delta R + \Delta R^T R - X^T\Delta X - \Delta X^T X)U^{-1}]U. \tag{2.10}$$

But for any symmetric matrix B,

$$2\|\mathrm{up}(B)\|_F^2 = \|B\|_F^2 - \frac{1}{2}(b_{11}^2 + b_{22}^2 + \cdots + b_{nn}^2) \le \|B\|_F^2.$$

Thus, from (2.10) we have

$$\begin{aligned}\|\widehat{\Delta U}\|_F &\le \frac{1}{\sqrt{2}}\|U^{-T}(R^T\Delta R + \Delta R^T R - X^T\Delta X - \Delta X^T X)U^{-1}\|_F\|U\|_2 \\ &\le \sqrt{2}\|U\|_2\|U^{-1}\|_2(\|RU^{-1}\|_2\|\Delta R\|_F + \|XU^{-1}\|_2\|\Delta X\|_F),\end{aligned}$$

which, combined with (2.2) and (2.3), gives

$$\|\widehat{\Delta U}\|_F \le \frac{\sqrt{2}\|U\|_2\|U^{-1}\|_2}{\sigma_n(\Gamma)}(\|\Delta R\|_F + \sqrt{1 - \sigma_n^2(\Gamma)}\|\Delta X\|_F),$$

resulting in the new perturbation bound for relative changes as

$$\frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} \le \sqrt{2}\,\frac{\|U^{-1}\|_2\|R\|_2}{\sigma_n(\Gamma)}\,\frac{\|\Delta R\|_F}{\|R\|_2} + \sqrt{2}\,\frac{\sqrt{1 - \sigma_n^2(\Gamma)}\|U^{-1}\|_2\|X\|_2}{\sigma_n(\Gamma)}\,\frac{\|\Delta X\|_F}{\|X\|_2}, \tag{2.11}$$

which leads to the condition numbers for the Cholesky downdating problem:

$$\beta_R \equiv \sqrt{2}\,\frac{\|U^{-1}\|_2\|R\|_2}{\sigma_n(\Gamma)}, \qquad \beta_X \equiv \sqrt{2}\frac{\sqrt{1 - \sigma_n^2(\Gamma)}\|U^{-1}\|_2\|X\|_2}{\sigma_n(\Gamma)}$$

for U with respect to relative changes in R and X, respectively. Notice from (2.1) $\beta_R > \beta_X$. So we can define a new overall condition number as

$$\beta_2 \equiv \beta_R = \sqrt{2}\,\frac{\|U^{-1}\|_2\|R\|_2}{\sigma_n(\Gamma)}. \tag{2.12}$$

Rewriting (2.11) as

$$\frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} \le \sqrt{2}\,\frac{\|U^{-1}\|_2\|R\|_2}{\sigma_n(\Gamma)}\left(\frac{\|\Delta R\|_F}{\|R\|_2} + \sqrt{1-\sigma_n^2(\Gamma)}\,\frac{\|X\|_2}{\|R\|_2}\,\frac{\|\Delta X\|_F}{\|X\|_2}\right),$$

and combining it with (2.5) and (2.7), gives Sun's bound

$$\frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} \le \sqrt{2}\,\frac{\kappa_2(U)}{\sigma_n^2(\Gamma)}\left(\frac{\|\Delta R\|_F}{\|R\|_2} + (1-\sigma_n^2(\Gamma))\frac{\|\Delta X\|_F}{\|X\|_2}\right). \tag{2.13}$$

We have seen the right-hand side of (2.11) is never worse than that of (2.13), so

$$\beta_2 \le \beta_1. \tag{2.14}$$

When $k = 1$, we have from (2.12) with $\sigma_n(\Gamma) = \sqrt{1-\|v\|_2^2}$ and (2.6) that

$$\beta_2 = \sqrt{2}\,\frac{\|U^{-1}\|_2\|R\|_2}{\sqrt{1-\|v\|_2^2}} \le \sqrt{2}\,\frac{\kappa_2(R)}{1-\|v\|_2^2} \le \sqrt{2}\,\frac{\kappa_2(R)(1+(\kappa_2(R)-1)\|v\|_2^2)}{1-\|v\|_2^2}.$$

Notice the most right-hand side is just β_0/c_n, thus

$$\beta_2 \le \beta_0/c_n \le \beta_0.$$

Although β_2 is a minor improvement on β_1, and is also an improvement on β_0 when $k = 1$, it is still not what we want. We can see this from the asymptotic behavior of these "condition numbers." The Cholesky factorization is unique, so as $X \to 0$, $U \to R$, and $X^T\Delta X \to 0$ in (2.9). Now for any upper triangular perturbation ΔR in R, $\Delta U \to \Delta R$, so the true condition number should approach unity. Here β_1, $\beta_2 \to \sqrt{2}\kappa_2(R)$. The next section shows how we can overcome this inadequacy.

3. New perturbation results. In section 2 we saw the key to deriving first-order perturbation bounds for U in the block Cholesky downdating problem is the equation (2.9). We will now analyze it in two new approaches. The two approaches have been used in the perturbation analyses of the Cholesky factorization, the QR factorization (see Chang, Paige, and Stewart [6, 7]), and LU factorization (see Chang [4] and Stewart [19]). The first approach, the refined matrix equation approach, gives a clear improvement on the previous results, while the second, the matrix–vector equation approach, gives a further improvement still, which leads to the true condition numbers for the block Cholesky downdating problem.

3.1. Refined matrix equation analysis. In the last section we used (2.9) to produce the matrix equation (2.10), and derived the bounds directly from this. We now look at this approach more closely.

Let $\mathcal{D}_n$ be the set of all $n \times n$ real positive definite diagonal matrices. For any $D = \mathrm{diag}(\delta_1, \ldots, \delta_n) \in \mathcal{D}_n$, let $U = D\bar{U}$. Note that for any matrix B we have $\mathrm{up}(BD^{-1}) = \mathrm{up}(B)D^{-1}$ and $\mathrm{up}(D^{-1}B) = D^{-1}\mathrm{up}(B)$.

First *with no restriction on* ΔR we have from (2.10)

$$\begin{aligned}\widehat{\Delta U} = &\{\mathrm{up}(U^{-T}R^T\Delta R\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}\Delta R^T R U^{-1})D\}\bar{U}\\ &- \{\mathrm{up}(U^{-T}X^T\Delta X\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}\Delta X^T X U^{-1})D\}\bar{U},\end{aligned}$$

so taking the F-norm gives

$$\begin{aligned}\|\widehat{\Delta U}\|_F \le &\|\mathrm{up}(U^{-T}R^T\Delta R\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}\Delta R^T R U^{-1})D\|_F\|\bar{U}\|_2\\ &+\|\mathrm{up}(U^{-T}X^T\Delta X\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}\Delta X^T X U^{-1})D\|_F\|\bar{U}\|_2.\end{aligned} \tag{3.1}$$

It is easy to show for any $B \in \mathcal{R}^{n\times n}$ (see Lemma 5.1 in [7])

$$\|\mathrm{up}(B) + D^{-1}\mathrm{up}(B^T)D\|_F \le \sqrt{1+\zeta_D^2}\|B\|_F, \tag{3.2}$$

where

$$\zeta_D = \max_{1\le i<j\le n}\{\delta_j/\delta_i\}. \tag{3.3}$$

Thus, from (3.1) we have

$$\begin{aligned}\|\widehat{\Delta U}\|_F &\le \sqrt{1+\zeta_D^2}(\|U^{-T}R^T\Delta R\bar{U}^{-1}\|_F + \|U^{-T}X^T\Delta X\bar{U}^{-1}\|_2)\|\bar{U}\|_2\\ &\le \sqrt{1+\zeta_D^2}\kappa_2(\bar{U})(\|RU^{-1}\|_2\|\Delta R\|_F + \|XU^{-1}\|_2\|\Delta X\|_F)\\ &= \sqrt{1+\zeta_D^2}\frac{\kappa_2(\bar{U})}{\sigma_n(\Gamma)}(\|\Delta R\|_F + \sqrt{1-\sigma_n^2(\Gamma)}\|\Delta X\|_F) \quad \text{(using (2.2), (2.3))}\end{aligned}$$

which is an elegant result in the changes alone. It leads to the following perturbation bound in terms of relative changes:

$$\begin{aligned}\frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} \le &\sqrt{1+\zeta_D^2}\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|R\|_2}{\|U\|_2}\frac{\|\Delta R\|_F}{\|R\|_2}\\ &+\sqrt{1+\zeta_D^2}\sqrt{1-\sigma_n^2(\Gamma)}\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|X\|_2}{\|U\|_2}\frac{\|\Delta X\|_F}{\|X\|_2}.\end{aligned} \tag{3.4}$$

Although here it would be simpler to just define an overall condition number, for later comparisons it is necessary for us to define the following two quantities as condition numbers for U with respect to relative changes in R and X, respectively (here subscript $_G$ refers to *general* ΔR, and later the subscript $_T$ will refer to *upper triangular* ΔR):

$$c_{RG}(R,X) \equiv \inf_{D\in\mathcal{D}_n} c_{RG}(R,X,D), \qquad c_X(R,X) \equiv \inf_{D\in\mathcal{D}_n} c_X(R,X,D), \tag{3.5}$$

where

$$c_{RG}(R,X,D) \equiv \sqrt{1+\zeta_D^2}\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|R\|_2}{\|U\|_2}, \tag{3.6}$$

$$c_X(R,X,D) \equiv \sqrt{1+\zeta_D^2}\sqrt{1-\sigma_n^2(\Gamma)}\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|X\|_2}{\|U\|_2}. \tag{3.7}$$

Then an overall condition number can be defined as

$$c_G(R,X) \equiv \inf_{D\in\mathcal{D}_n} c_G(R,X,D), \tag{3.8}$$

where

$$c_G(R,X,D) \equiv \max\{c_{RG}(R,X,D), c_X(R,X,D)\} = c_{RG}(R,X,D).$$

Obviously we have

$$c_G(R,X) = c_{RG}(R,X) \geq c_X(R,X). \tag{3.9}$$

Thus, with these, we have from (3.4) that

$$\begin{aligned} \frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} &\leq c_{RG}(R,X)\frac{\|\Delta R\|_2}{\|R\|_2} + c_X(R,X)\frac{\|\Delta X\|_F}{\|X\|_2} \\ &\leq c_G(R,X)\left(\frac{\|\Delta R\|_F}{\|R\|_2} + \frac{\|\Delta X\|_F}{\|X\|_2}\right). \end{aligned} \tag{3.10}$$

Clearly if we take $D = I_n$, (3.4) will become (2.11), and

$$\begin{aligned} c_{RG}(R,X) \leq c_{RG}(R,X,I_n) = \beta_R, \qquad c_X(R,X) \leq c_X(R,X,I_n) = \beta_X, \\ c_G(R,X) \leq c_G(R,X,I_n) = \beta_2. \end{aligned} \tag{3.11}$$

It is not difficult to give an example to show β_2 can be arbitrarily larger than $c_G(R,X)$, as can be seen from the following asymptotic behavior.

If $X \to 0$ we saw $U \to R$ and $\sigma_n(\Gamma) \to 1$, so

$$c_G(R,X,D) \to \sqrt{1+\zeta_D^2}\kappa_2(D^{-1}R).$$

It is shown in [7, sect. 5.1, (5.14)] that with an appropriate choice of D, $\sqrt{1+\zeta_D^2}\kappa_2(D^{-1}R)$ has a bound which is a function of n only, if R was found using the standard pivoting strategy in the Cholesky factorization, and in this case, we see the condition number $c_G(R,X)$ of the problem here is bounded independently of $\kappa_2(R)$ as $X \to 0$, for *general* ΔR. At the end of this section we give an even stronger result when $X \to 0$ for the case of upper triangular ΔR. Note in the case here that β_2 in (2.12) can be made as large as we like, and thus arbitrarily larger than $c_G(R,X)$.

In the case where ΔR is *upper triangular*, we can refine the analysis further. From (2.10) we have

$$\begin{aligned} \widehat{\Delta U} = [&\mathrm{up}(U^{-T}R^T\Delta R U^{-1} + U^{-T}\Delta R^T R U^{-1}) \\ &- \mathrm{up}(U^{-T}X^T\Delta X U^{-1} + U^{-T}\Delta X^T X U^{-1})]U. \end{aligned} \tag{3.12}$$

Notice with (1.3) and (1.4)

$$\begin{aligned} & U^{-T}R^T\Delta R U^{-1} + U^{-T}\Delta R^T R U^{-1} \\ &= [\mathrm{slt}(U^{-T}R^T) + \mathrm{diag}(U^{-T}R^T)]\Delta R U^{-1} \\ &\quad + U^{-T}\Delta R^T[\mathrm{sut}(RU^{-1}) + \mathrm{diag}(RU^{-1})] \\ &= \mathrm{diag}(U^{-T}R^T)\cdot\Delta R U^{-1} + U^{-T}\Delta R^T\cdot\mathrm{diag}(RU^{-1}) \\ &\quad + \mathrm{slt}(U^{-T}R^T)\cdot\Delta R U^{-1} + U^{-T}\Delta R^T\cdot\mathrm{sut}(RU^{-1}). \end{aligned} \tag{3.13}$$

But for any *upper* triangular matrix T we have

$$\mathrm{up}(T) + \mathrm{up}(T^T) = T,$$

so that if we define $T \equiv \mathrm{diag}(U^{-T}R^T)\cdot\Delta RU^{-1}$, then

$$\begin{aligned}(3.14)\qquad &\mathrm{up}[\mathrm{diag}(U^{-T}R^T)\cdot\Delta RU^{-1} + U^{-T}\Delta R^T\cdot\mathrm{diag}(RU^{-1})]\\ &= \mathrm{diag}(U^{-T}R^T)\cdot\Delta RU^{-1}.\end{aligned}$$

Thus, from (3.12), (3.13), and (3.14) we obtain

$$\begin{aligned}(3.15)\qquad &\widehat{\Delta U}\\ &= \mathrm{diag}(U^{-T}R^T)\cdot\Delta R + \{\mathrm{up}[\mathrm{slt}(U^{-T}R^T)\cdot\Delta RU^{-1} + U^{-T}\Delta R^T\cdot\mathrm{sut}(RU^{-1})]\\ &\quad - \mathrm{up}(U^{-T}X^T\Delta XU^{-1} + U^{-T}\Delta X^TXU^{-1})\}U.\end{aligned}$$

As before, let $U = D\bar{U}$, where $D = \mathrm{diag}(\delta_1,\ldots,\delta_n)\in\mathcal{D}_n$. From (3.15) it follows that

$$\begin{aligned}\|\widehat{\Delta U}\|_F &\le \|\mathrm{diag}(U^{-T}R^T)\|_2\|\Delta R\|_F\\ &\quad + \|\mathrm{up}[\mathrm{slt}(U^{-T}R^T)\cdot\Delta R\bar{U}^{-1}] + D^{-1}\mathrm{up}[\bar{U}^{-T}\Delta R^T\cdot\mathrm{sut}(RU^{-1})]D\|_F\|\bar{U}\|_2\\ &\quad + \|\mathrm{up}(U^{-T}X^T\Delta X\bar{U}^{-1}) + D^{-1}\mathrm{up}(\bar{U}^{-T}\Delta X^TXU^{-1})D\|_F\|\bar{U}\|_2.\end{aligned}$$

Then, applying (3.2) to this, we get the following perturbation bound:

$$\begin{aligned}\frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} &\le \big(\|\mathrm{diag}(RU^{-1})\|_2 + \sqrt{1+\zeta_D^2}\kappa_2(D^{-1}U)\|\mathrm{sut}(RU^{-1})\|_2\big)\frac{\|R\|_2}{\|U\|_2}\frac{\|\Delta R\|_F}{\|R\|_2}\\ &\quad + \sqrt{1+\zeta_D^2}\kappa_2(D^{-1}U)\|XU^{-1}\|_2\frac{\|X\|_2}{\|U\|_2}\frac{\|\Delta X\|_F}{\|X\|_2}.\end{aligned}\qquad(3.16)$$

Comparing (3.16) with (3.4) and noticing (2.3), we see (trivially) the sensitivity of U with respect to changes in X does not change, so $c_X(R,X)$ defined in (3.5) can still be regarded as a condition number for U with respect to changes in X. But we now need to define a new condition number for U with respect to *upper triangular* changes in R, that is (subscript $_T$ indicates upper triangular ΔR),

$$c_{RT}(R,X) \equiv \inf_{D\in\mathcal{D}_n} c_{RT}(R,X,D),$$

where

$$(3.17)\quad c_{RT}(R,X,D) \equiv \left(\|\mathrm{diag}(RU^{-1})\|_2 + \sqrt{1+\zeta_D^2}\kappa_2(D^{-1}U)\|\mathrm{sut}(RU^{-1})\|_2\right)\frac{\|R\|_2}{\|U\|_2}.$$

Thus, an overall condition number can be defined as

$$c_T(R,X) = \inf_{D\in\mathcal{D}_n} c_T(R,X,D),$$

where

$$c_T(R,X,D) = \max\{c_{RT}(R,X,D), c_X(R,X,D)\}.$$

Obviously we have

$$(3.18)\qquad c_T(R,X) = \max\{c_{RT}(R,X), c_X(R,X)\}.$$

With these we have from (3.16) that

$$\begin{aligned}(3.19)\qquad \frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} &\le c_{RT}(R,X)\frac{\|\Delta R\|_F}{\|R\|_2} + c_X(R,X)\frac{\|\Delta X\|_F}{\|X\|_2}\\ &\le c_T(R,X)\left(\frac{\|\Delta R\|_F}{\|R\|_2} + \frac{\|\Delta X\|_F}{\|X\|_2}\right).\end{aligned}$$

What is the relationship between $c_T(R,X)$ and $c_G(R,X) = c_{RG}(R,X)$? For any $n \times n$ upper triangular matrix $T = (t_{ij})$, observe the following two facts:

1) t_{ii}, $i = 1, 2, \ldots, n$ are the eigenvalues of T, so that

$$|t_{ii}| \le \|T\|_2,$$

which gives

$$\|\mathrm{diag}(T)\|_2 \le \|T\|_2.$$

2) $$\|\mathrm{sut}(T)\|_2 \le \|\mathrm{sut}(T)\|_F \le \|T\|_F \le \sqrt{n}\|T\|_2.$$

(Note: In fact we can prove a slightly sharper inequality $\|\mathrm{sut}(T)\|_2 \le \sqrt{n-1}\|T\|_2$).

Therefore,

$$\begin{aligned}
c_{RT}(R,X,D) &= \big(\|\mathrm{diag}(RU^{-1})\|_2 + \sqrt{1+\zeta_D^2}\kappa_2(D^{-1}U)\|\mathrm{sut}(RU^{-1})\|_2\big)\frac{\|R\|_2}{\|U\|_2} \\
&\le \big(\|RU^{-1}\|_2 + \sqrt{n}\sqrt{1+\zeta_D^2}\kappa_2(D^{-1}U)\|RU^{-1}\|_2\big)\frac{\|R\|_2}{\|U\|_2} \\
&< (1+\sqrt{n})\sqrt{1+\zeta_D^2}\kappa_2(D^{-1}U)\|RU^{-1}\|_2\frac{\|R\|_2}{\|U\|_2} \\
&= (1+\sqrt{n})\sqrt{1+\zeta_D^2}\frac{\kappa_2(D^{-1}U)}{\sigma_n(\Gamma)}\frac{\|R\|_2}{\|U\|_2} \quad \text{(using (2.2))} \\
&= (1+\sqrt{n})c_{RG}(R,X,D),
\end{aligned}$$

so that

$$c_{RT}(R,X) \le (1+\sqrt{n})c_{RG}(R,X).$$

Thus, we have from (3.9) and (3.18)

$$c_T(R,X) \le (1+\sqrt{n})c_G(R,X). \tag{3.20}$$

On the other hand, $c_T(R,X)$ can be arbitrarily smaller than $c_G(R,X)$. This can be seen from the asymptotic behavior, which is important in its own right. As $X \to 0$, since $U \to R$, $\sigma_n(\Gamma) \to 1$, and $RU^{-1} \to I_n$, we have

$$c_T(R,X,I_n) \to 1,$$

so for upper triangular changes in R, whether pivoting was used in finding R or not,

$$c_T(R,X) \to 1.$$

Thus, when $X \to 0$, the bound in (3.19) reflects the true sensitivity of the problem. For the case of general ΔR, if we do not use pivoting it is straightforward to make $c_G(R,X)$ in (3.8) arbitrarily large even with $X = 0$; see (3.6).

3.2. Matrix–vector equation analysis. In the last section, based on the structure of ΔR, we gave two perturbation bounds using the so-called refined matrix equation approach. Also based on the structure of ΔR, we can now obtain provably sharp

but less intuitive results by viewing the matrix equation (2.9) as a large matrix–vector equation. For any matrix $C \equiv (c_{ij}) \equiv [c_1, \ldots, c_n] \in \mathcal{R}^{n\times n}$, denote by $c_j^{(i)}$ the vector of the first i elements of c_j. With this, we define ("u" denotes "upper")

$$\mathrm{uvec}(C) \equiv \begin{pmatrix} c_1^{(1)} \\ c_2^{(2)} \\ \cdot \\ c_n^{(n)} \end{pmatrix}.$$

It is the vector formed by stacking the columns of the upper triangular part of C into one long vector.

First assume ΔR is a *general* real $n \times n$ matrix. It is easy to show (2.9) can be rewritten into the following matrix–vector form (cf. [7]):

$$W_U \mathrm{uvec}(\widehat{\Delta U}) = Z_R \mathrm{vec}(\Delta R) - Y_X \mathrm{vec}(\Delta X), \tag{3.21}$$

where $W_U \in \mathcal{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ is

$$\begin{bmatrix}
u_{11} & & & & & & & & & & & \\
u_{12} & u_{11} & & & & & & & & & & \\
 & u_{12} & u_{22} & & & & & & & & & \\
u_{13} & & & u_{11} & & & & & & & & \\
 & u_{13} & u_{23} & u_{12} & u_{22} & & & & & & & \\
 & & & u_{13} & u_{23} & u_{33} & & & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & & \\
u_{1n} & & & & & & & u_{11} & & & & \\
 & u_{1n} & u_{2n} & & & & & u_{12} & u_{22} & & & \\
 & & & u_{1n} & u_{2n} & u_{3n} & & u_{13} & u_{23} & u_{33} & & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \\
 & & & & & & & u_{1n} & u_{2n} & u_{3n} & \cdot & u_{nn}
\end{bmatrix},$$

$Z_R \in \mathcal{R}^{\frac{n(n+1)}{2}\times n^2}$ is

$$\left[\begin{array}{cccc|cccc|c|cccc}
r_{11} & & & & & & & & & & & \\
r_{12} & r_{22} & & & r_{11} & & & & & & & \\
 & & & & r_{12} & r_{22} & & & & & & \\
\cdot & \cdot & \cdot & & \cdot & \cdot & \cdot & & \cdot & & & \\
r_{1n} & r_{2n} & \cdot & r_{nn} & & & & & & r_{11} & & \\
 & & & & r_{1n} & r_{2n} & \cdot & r_{nn} & & r_{12} & r_{22} & \\
 & & & & & & & & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & & & r_{1n} & r_{2n} & \cdot & r_{nn}
\end{array}\right],$$

and $Y_X \in \mathcal{R}^{\frac{n(n+1)}{2}\times kn}$ is

$$\left[\begin{array}{cccc|cccc|c|cccc}
x_{11} & x_{21} & \cdot & x_{k1} & & & & & & & & \\
x_{12} & x_{22} & \cdot & x_{k2} & x_{11} & x_{21} & \cdot & x_{k1} & & & & \\
 & & & & x_{12} & x_{22} & \cdot & x_{k2} & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & \\
x_{1n} & x_{2n} & \cdot & x_{kn} & & & & & & x_{11} & x_{21} & \cdot & x_{k1} \\
 & & & & x_{1n} & x_{2n} & \cdot & x_{kn} & & x_{12} & x_{22} & \cdot & x_{kn} \\
 & & & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & & & x_{1n} & x_{2n} & \cdot & x_{nn}
\end{array}\right].$$

Since U is nonsingular, W_U is also, and from (3.21)

$$\text{uvec}(\widehat{\Delta U}) = W_U^{-1} Z_R \text{vec}(\Delta R) - W_U^{-1} Y_X \text{vec}(\Delta X), \tag{3.22}$$

so taking the 2-norm gives

$$\|\widehat{\Delta U}\|_F \le \|W_U^{-1} Z_R\|_2 \|\Delta R\|_F + \|W_U^{-1} Y_X\|_2 \|\Delta X\|_F,$$

resulting in the following perturbation bound:

$$\begin{aligned} \frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} &\le \kappa_{RG}(R,X) \frac{\|\Delta R\|_F}{\|R\|_2} + \kappa_X(R,X) \frac{\|\Delta X\|_F}{\|X\|_2} \\ &\le \kappa_{CDG}(R,X) \left(\frac{\|\Delta R\|_F}{\|R\|_2} + \frac{\|\Delta X\|_F}{\|X\|_2} \right), \end{aligned} \tag{3.23}$$

where

$$\kappa_{RG}(R,X) \equiv \frac{\|W_U^{-1} Z_R\|_2 \|R\|_2}{\|U\|_2}, \qquad \kappa_X(R,X) \equiv \frac{\|W_U^{-1} Y_X\|_2 \|X\|_2}{\|U\|_2}, \tag{3.24}$$

$$\kappa_{CDG}(R,X) \equiv \max\{\kappa_{RG}(R,X), \kappa_X(R,X)\}. \tag{3.25}$$

Now we would like to show

$$\kappa_{RG}(R,X) \le c_{RG}(R,X), \qquad \kappa_X(R,X) \le c_X(R,X). \tag{3.26}$$

Before showing this, we will prove a more general result. Suppose from (2.9) we are able to obtain a perturbation bound of the form

$$\frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} \le \alpha_R \frac{\|\Delta R\|_F}{\|R\|_2} + \alpha_X \frac{\|\Delta X\|_F}{\|X\|_2}, \tag{3.27}$$

where α_R and α_X, two functions of R and X, are other measures of the sensitivity of the Cholesky downdating problem with respect to changes in R and X. Let $\Delta X = 0$. Then from (3.22) and (3.27) we have

$$\frac{\|W_U^{-1} Z_R \text{vec}(\Delta R)\|_2}{\|U\|_2} \le \alpha_R \frac{\|\Delta R\|_F}{\|R\|_2}.$$

Notice ΔR can be any (sufficiently small) $n \times n$ real matrix, so we must have

$$\|W_U^{-1} Z_R\|_2 \le \alpha_R \frac{\|U\|_2}{\|R\|_2},$$

which gives

$$\kappa_{RG}(R,X) \le \alpha_R.$$

Similarly, we can show

$$\kappa_X(R,X) \le \alpha_X.$$

Notice since (3.10) is a particular case of (3.27), (3.26) follows. Thus, we have from (3.9) and (3.25)

$$\kappa_{CDG}(R,X) \le c_G(R,X). \tag{3.28}$$

The above analysis shows for general ΔR, $\kappa_{RG}(R,X)$ and $\kappa_X(R,X)$ are optimal measures of the sensitivity of U with respect to changes in R and X, respectively, and thus the bound (3.23) is optimal. So we propose $\kappa_{RG}(R,X)$ and $\kappa_X(R,X)$ as the *true* condition numbers for U with respect to general changes in R and X, respectively, and $\kappa_{CDG}(R,X)$ as the *true* overall condition number of the problem in this case.

It is easy to observe that if $X \to 0$, $\kappa_{CDG}(R,X) \to \|W_R^{-1}Z_R\|_2$, where W_R is just W_U with each entry u_{ij} replaced by r_{ij}. If R was found using the standard pivoting strategy in the Cholesky factorization, then $\|W_R^{-1}Z_R\|_2$ has a bound which is a function of n alone (see [5] for a proof). So in this case our condition number $\kappa_{CDG}(R,X)$ also has a bound which is a function of n alone as $X \to 0$.

Remark 1. Our numerical experiments suggest $c_G(R,X)$ is usually a good approximation to $\kappa_{CDG}(R,X)$. But the following example shows $c_G(R,X)$ can sometimes be arbitrarily larger than $\kappa_{CDG}(R,X)$:

$$R = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \epsilon^3 & 0 \\ 0 & 0 & 0 & \epsilon^2 \end{pmatrix}, \quad X = \begin{pmatrix} \sqrt{3} & 2/\sqrt{3} & 0 & 0 \\ 0 & \sqrt{2/3-\epsilon^2} & 0 & 0 \end{pmatrix}, \quad U = \mathrm{diag}(1,\epsilon,\epsilon^3,\epsilon^2),$$

where ϵ is a small positive number. It is not difficult to show

$$c_G(R,X) = O\left(\frac{1}{\epsilon^2}\right), \qquad \kappa_{CDG}(R,X) = O\left(\frac{1}{\epsilon}\right).$$

But $c_G(R,X)$ has an advantage over $\kappa_{CDG}(R,X)$ — it can be quite easy to estimate — all we need do is choose a suitable D in $c_G(R,X,D)$. We consider how to do this in the next section. In contrast $\kappa_{CDG}(R,X)$ is, as far as we can see, unreasonably expensive to compute or estimate.

Now we consider the case where ΔR is *upper triangular*. Equation (2.9) can now be rewritten as the following matrix–vector form:

$$W_U \mathrm{uvec}(\widehat{\Delta U}) = W_R \mathrm{uvec}(\Delta R) - Y_X \mathrm{vec}(\Delta X), \tag{3.29}$$

where $W_U \in \mathcal{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ and $Y_X \in \mathcal{R}^{\frac{n(n+1)}{2}\times kn}$ are defined as before, and $W_R \in \mathcal{R}^{\frac{n(n+1)}{2}\times\frac{n(n+1)}{2}}$ is just W_U with each entry u_{ij} replaced by r_{ij}. Since U is nonsingular, W_U is also, and from (3.29)

$$\mathrm{uvec}(\widehat{\Delta U}) = W_U^{-1}W_R \mathrm{uvec}(\Delta R) - W_U^{-1}Y_X \mathrm{vec}(\Delta X),$$

so taking the 2-norm gives

$$\|\widehat{\Delta U}\|_F \le \|W_U^{-1}W_R\|_2 \|\Delta R\|_F + \|W_U^{-1}Y_X\|_2 \|\Delta X\|_F,$$

which leads to the following perturbation bound:

$$\begin{aligned} \frac{\|\widehat{\Delta U}\|_F}{\|U\|_2} &\le \kappa_{RT}(R,X)\frac{\|\Delta R\|_F}{\|R\|_2} + \kappa_X(R,X)\frac{\|\Delta X\|_F}{\|X\|_2} \\ &\le \kappa_{CDT}(R,X)\left(\frac{\|\Delta R\|_F}{\|R\|_2} + \frac{\|\Delta X\|_F}{\|X\|_2}\right), \end{aligned} \tag{3.30}$$

where

$$\kappa_{RT}(R,X) \equiv \frac{\|W_U^{-1}W_R\|_2\|R\|_2}{\|U\|_2}, \qquad \kappa_X(R,X) \equiv \frac{\|W_U^{-1}Y_X\|_2\|X\|_2}{\|U\|_2},$$

$$\kappa_{CDT}(R,X) \equiv \max\{\kappa_{RT}(R,X), \kappa_X(R,X)\}. \tag{3.31}$$

Note $\kappa_X(R,X)$ is the same as that defined in (3.24).

As before, we can show that for the case where ΔR is upper triangular, $\kappa_{RT}(R,X)$ and $\kappa_X(R,X)$ are optimal measures of the sensitivity of U with respect to changes in R and X, respectively, and thus the bound (3.30) is optimal. In particular, we have

$$\kappa_{RT}(R,X) \le c_{RG}(R,X), c_{RT}(R,X), \kappa_{RG}(R,X); \qquad \kappa_X(R,X) \le c_X(R,X).$$

In fact $\kappa_{RT}(R,X) \le \kappa_{RG}(R,X)$ can also be proved directly by the fact that the columns of W_R form a proper subset of the columns of Z_R, and the second inequality has been proved before. Thus, we have from (3.9), (3.18), (3.25), and (3.31)

$$\kappa_{CDT}(R,X) \le c_G(R,X), c_T(R,X), \kappa_{CDG}(R,X). \tag{3.32}$$

By the above analysis, we propose $\kappa_{RT}(R,X)$ and $\kappa_X(R,X)$ as the *true* condition numbers for U with respect to changes in R and X, respectively, and $\kappa_{CDT}(R,X)$ as the *true* overall condition number, in the case that ΔR is upper triangular.

If as well $X \to 0$, then since $U \to R$, $W_U^{-1}W_R \to I_{\frac{n(n+1)}{2}}$, and $\kappa_{CDT}(R,X) \to 1$. So in this case the Cholesky downdating problem becomes very well conditioned no matter how ill conditioned R or U is.

Remark 2. Numerical experiments also suggest $c_T(R,X)$ is usually a good approximation to $\kappa_{CDT}(R,X)$. But sometimes $c_T(R,X)$ can be arbitrarily larger than $\kappa_{CDT}(R,X)$. This can also be seen from the example in Remark 1. In fact, it is not difficult to obtain

$$c_T(R,X) = O\left(\frac{1}{\epsilon^2}\right), \qquad \kappa_{CDT}(R,X) = O\left(\frac{1}{\epsilon}\right).$$

Like $\kappa_{CDG}(R,X)$, $\kappa_{CDT}(R,X)$ is difficult to compute or estimate. But $c_T(R,X)$ is easy to estimate, which is discussed in the next section.

4. Numerical tests and condition estimators. In section 3 we presented new first-order perturbation bounds for the downdated Cholesky factor U using first the refined matrix equation approach, and then the matrix–vector equation approach. We defined $\kappa_{CDG}(R,X)$ for general ΔR, and $\kappa_{CDT}(R,X)$ for upper triangular ΔR, as the true overall condition numbers of the problem. Also we gave two corresponding practical but weaker condition numbers $c_G(R,X)$ and $c_T(R,X)$ for the two ΔR cases.

We would like to choose D such that $c_G(R,X,D)$ and $c_T(R,X,D)$ are good approximations to $c_G(R,X)$ and $c_T(R,X)$, respectively. We see from (3.6), (3.7), and (3.17) that we want to find D such that $\sqrt{1+\zeta_D^2}\kappa_2(D^{-1}U)$ approximates its infimum. By a well-known result of van der Sluis [21], $\kappa_2(D^{-1}U)$ will be nearly minimal when the rows of $D^{-1}U$ are equilibrated. But this could lead to a large ζ_D. So a reasonable compromise is to choose D to equilibrate U as far as possible while keeping $\zeta_D \le 1$. Specifically, take $\zeta_1 = \sqrt{\sum_{j=1}^n u_{1j}^2}$, $\zeta_i = \sqrt{\sum_{j=i}^n u_{ij}^2}$ if $\sqrt{\sum_{j=i}^n u_{ij}^2} \le \zeta_{i-1}$ otherwise $\zeta_i = \zeta_{i-1}$, for $i = 2, \ldots, n$. Then we use a standard condition estimator to estimate $\kappa_2(D^{-1}U)$ in $O(n^2)$ operations.

Notice from (2.4) we have $\sigma_n(\Gamma) = \sqrt{1 - \|XR^{-1}\|_2^2}$. Usually k, the number of rows of X, is much smaller than n, so $\sigma_n(\Gamma)$ can be computed in $O(n^2)$. If k is not much smaller than n, then we use a standard norm estimator to estimate $\|XR^{-1}\|_2$ in $O(n^2)$. Similarly $\|U\|_2$ and $\|R\|_2$ can be estimated in $O(n^2)$. So finally $c_G(R,X,D)$ can be estimated in $O(n^2)$. Estimating $c_T(R,X,D)$ is not as easy as estimating $c_G(R,X,D)$. The part $\|\mathrm{diag}(RU^{-1})\|_2$ in $c_{RT}(R,X,D)$ can easily be computed in $O(n)$, since $\mathrm{diag}(RU^{-1}) = \mathrm{diag}(r_{11}/u_{11}, \ldots, r_{nn}/u_{nn})$. The part $\|\mathrm{sut}(RU^{-1})\|_2$ in

TABLE 4.1.

τ	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
$\|XR^{-1}\|_2$	0.99999	0.999	0.9	0.09	0.0009	0.000009
β_1	2.25e+10	2.25e+07	2.60e+04	2.72e+03	2.69e+03	2.69e+03
β_2	1.01e+08	1.01e+06	1.14e+04	2.71e+03	2.69e+03	2.69e+03
$c_G(R,X,D)$	3.60e+03	3.61e+02	3.79e+01	1.79e+01	1.78e+01	1.78e+01
$\kappa_{CDG}(R,X)$	1.66e+03	1.66e+02	1.71e+01	8.42e+00	8.41e+00	8.41e+00
$c_T(R,X,D)$	2.12e+03	2.12e+02	1.79e+01	1.07e+00	1.00e+00	1.00e+00
$\kappa_{CDT}(R,X)$	2.43e+02	2.43e+01	2.44e+00	1.01e+00	1.00e+00	1.00e+00

$c_{RT}(R,X,D)$ can roughly be estimated in $O(n^2)$, based on

$$\frac{1}{\sqrt{n-1}}\|\mathrm{sut}(RU^{-1})\|_F \le \|\mathrm{sut}(RU^{-1})\|_2 \le \|\mathrm{sut}(RU^{-1})\|_F,$$
$$\|\mathrm{sut}(RU^{-1})\|_F = \sqrt{\|RU^{-1}\|_F^2 - \|\mathrm{diag}(RU^{-1})\|_F^2},$$

and the fact that $\|RU^{-1}\|_F$ can be estimated by a standard norm estimator in $O(n^2)$. The value of $\|XU^{-1}\|_2$ in $c_X(R,X,D)$ can be calculated (if $k \ll n$) or estimated by a standard estimator in $O(n^2)$. All the remaining values $\|R\|_2$, $\|X\|_2$, and $\|U\|_2$ can also be estimated by a standard norm estimator in $O(n^2)$. Hence $c_{RT}(R,X,D)$, $c_X(R,X,D)$, and thus $c_T(R,X,D)$ can be estimated in $O(n^2)$. For standard condition estimators and norm estimators, see Chapter 14 of [14].

The relationships among the various overall condition numbers for the Cholesky downdating problem presented in sections 2 and 3 are as follows (see (2.14), (3.11), (3.28), (3.32), and (3.20)):

$$\beta_1 \ge \beta_2 \ge c_G(R,X) \ge \kappa_{CDG}(R,X) \ge \kappa_{CDT}(R,X),$$
$$(1+\sqrt{n})c_G(R,X) \ge c_T(R,X) \ge \kappa_{CDT}(R,X).$$

Now we give one numerical example to illustrate these. The example, quoted from Sun [20], is as follows:

$$R = \mathrm{diag}(1,s,s^2,s^3,s^4)\begin{pmatrix} 1 & -c & -c & -c & -c \\ 0 & 1 & -c & -c & -c \\ 0 & 0 & 1 & -c & -c \\ 0 & 0 & 0 & 1 & -c \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad X^T = \tau\begin{pmatrix} 0.240 \\ -0.899 \\ 0.899 \\ 1.560 \\ 2.390 \end{pmatrix},$$

where $c = 0.95$, $s = \sqrt{1-c^2}$. The results obtained using MATLAB are shown in Table 4.1 for various values of τ:

$$\tau_1 = 1.004015006005433\mathrm{e}-2, \quad \tau_2 = 1.003021021209640\mathrm{e}-2,$$
$$\tau_3 = 9.036225416303058\mathrm{e}-3,$$

and $\tau_4 = \tau_3 \cdot \mathrm{e}-01$, $\tau_5 = \tau_3 \cdot \mathrm{e}-03$, $\tau_6 = \tau_3 \cdot \mathrm{e}-5$.

Note in Table 4.1 how β_1 and β_2 can be far worse than the true condition numbers $\kappa_{CDG}(R,X)$ and $\kappa_{CDT}(R,X)$, although β_2 is not as bad as β_1. Also we observe that $c_G(R,X,D)$ and $c_T(R,X,D)$ are very good approximations to $\kappa_{CDG}(R,X)$ and $\kappa_{CDT}(R,X)$, respectively. When X become small, all of the condition numbers decrease. The asymptotic behavior of $c_G(R,X,D)$, $c_T(R,X,D)$, $\kappa_{CDG}(R,X)$, and $\kappa_{CDT}(R,X)$ coincides with our theoretical results — when $X \to 0$, $c_G(R,X)$ and $\kappa_{CDG}(R,X)$ will be bounded in terms of n since here R corresponds to the Cholesky factor of a correctly pivoted A, and $c_T(R,X), \kappa_{CDT}(R,X) \to 1$.

Acknowledgment. We would like to thank Ji-Guang Sun for suggesting to us that the approach described in (a draft version of) [6] might apply to the Cholesky downdating problem.

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] A. Björck, H. Park, and L. Eldén, *Accurate downdating of least squares solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 549–568.

[3] A. W. Bojanczyk, R. P. Brent, P. Van Dooren, and F. R. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 210–221.

[4] X.-W. Chang, *Perturbation Analysis of Some Matrix Factorizations*, Ph.D. thesis, School of Computer Science, McGill University, Montreal, 1997.

[5] X.-W. Chang and C. C. Paige, *A Perturbation Analysis for R in the QR Factorization*, Tech. report SOCS-95.7, School of Computer Science, McGill University, Montreal, 1995.

[6] X.-W. Chang, C. C. Paige, and G. W. Stewart, *New perturbation analyses for the Cholesky factorization*, IMA J. Numer. Anal., 16 (1996), pp. 457–484.

[7] X.-W. Chang, C. C. Paige, and G. W. Stewart, *Perturbation analyses for the QR factorization*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 775–791.

[8] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, *LINPACK User's Guide*, SIAM, Philadelphia, PA, 1979.

[9] L. Eldén and H. Park, *Block downdating of least squares solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1018–1034.

[10] L. Eldén and H. Park, *Perturbation analysis for block downdating of a Cholesky decomposition*, Numer. Math., 68 (1994), pp. 457–467.

[11] P. E. Gill, G. H. Golub, W. Murray, and M. A. Saunders, *Methods for modifying matrix factorizations*, Math. Comp., 28 (1974), pp. 505–535.

[12] G. H. Golub and G. P. Styan, *Numerical computations for univariate linear models*, J. Statist. Comput. Simul., 2 (1973), pp. 253–272.

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

[14] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.

[15] C.-T. Pan, *A perturbation analysis of the problem of downdating a Cholesky factorization*, Linear Algebra Appl., 183 (1993), pp. 103–116.

[16] C.-T. Pan and R. Plemmons, *Least squares modification with inverse factorizations: parallel implications*, J. Comput. Appl. Math., 27 (1989), pp. 109–127.

[17] M. A. Saunders, *Large-Scale Linear Programming Using the Cholesky Factorization*, Tech. report CS252, Computer Science Department, Stanford University, Standord, CA, 1972.

[18] G. W. Stewart, *The effects of rounding error on an algorithm for downdating a Cholesky factorization*, J. Inst. Math. Appl., 23 (1979), pp. 203–213.

[19] G. W. Stewart, *On the Perturbation of LU and Cholesky Factors*, Tech. report CS-TR-3535 UMIACS-TR-95-93, Department of Computer Science, University of Maryland, College Park, MD, 1995.

[20] J.-G. Sun, *Perturbation analysis of the Cholesky downdating and QR updating problems*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 760–775.

[21] A. van der Sluis, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 444–467, April 1998

011

SINGLE-INPUT EIGENVALUE ASSIGNMENT ALGORITHMS: A CLOSE LOOK*

MARK ARNOLD† AND BISWA NATH DATTA‡

Abstract. A close look is taken at the single-input eigenvalue assignment methods. Several previously known backward stable QR algorithms are tied together in a common framework of which each is a special case, and their connection to an explicit expression for the feedback vector is exposed. A simple new algorithm is presented and its backward stability is established by round-off error analysis. The differences between this new algorithm and the other QR algorithms are discussed. Also, the round-off error analysis of a simple recursive algorithm for the problem [B. N. Datta, *IEEE Trans. Automat. Control*, AC-32 (1987), pp. 414–417] is presented. The analysis shows that the latter is reliable, and the reliability can be determined during the execution of the algorithm rather cheaply. Finally, some numerical experiments comparing some of the methods are reported.

Key words. eigenvalue assignment, unified approach

AMS subject classifications. 15, 65F, 65G, 93B

PII. S0895479895294885

1. Introduction.

Given a controllable pair of matrices (A, B) and a set $\Omega = \{\lambda_1, ..., \lambda_n\}$, closed under complex conjugation, the well-known eigenvalue assignment problem in control theory is the problem of finding a matrix F such that $A + BF$ has the spectrum Ω (see Chen [11], Kailath [24], Szidarovszky and Bahill [38]).

Because of its importance, the problem has been very well studied in both mathematics and control literatures. Many methods exist: single-input and multi-input (Arnold and Datta [3], Bhattacharyya and DeSouza [4], Bru, Mas, and Urbano [8], Bru, Cerdan, and Urbano [9], Datta [17], Miminis and Paige [27, 29], Patel and Misra [32], Petkov, Christov, and Konstantinov [33, 34], Tsui [37], Varga [40]); robust eigenvalue assignment (Kautsky, Nichols, and Van Dooren [25]); partial eigenvalue assignment (Datta and Saad [23], Saad [35]); parallel algorithms (Arnold [2], Bru, Cerdan, Fernandez de Cordoba, and Urbano [10], Datta [20], Datta [19], Coutinho, Bhaya, and Datta [15], Datta and Datta [18],); and methods for second-order systems (Datta, Elhay, and Ram [22], Chu and Datta [12]). The backward stability of some of these algorithms have been established by round-off error analysis (Cox and Moss [13, 14], Miminis and Paige [29]).

We take another look at the single-input methods in this paper.

In theory all the single-input algorithms produce the same solution (see Wonham [42]). It is therefore natural to explore the relationships between these methods. We relate the QR methods of Miminis and Paige [27], Patel and Misra [32], and Petkov, Christov, and Konstantinov [34] under one umbrella and then relate the recursive algorithm of Datta [17] to these results. Specifically, we prove a result that shows that all these methods are connected by a simple property of QR iteration and the

*Received by the editors November 17, 1995; accepted for publication (in revised form) by P. Van Dooren April 3, 1997.

http://www.siam.org/journals/simax/19-2/29488.html

†Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701 (arnold@mysong.uark.edu).

‡Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL 60115 (dattab@math.niu.edu). The research of this author was supported by an NSF grant under contract DMS-9212629.

explicit closed form solution of the single-input eigenvalue assignment problem that can be obtained easily from the recursive algorithm.

These results do not seem to have appeared in the literature before. The relationship allows us to present the QR algorithms in a unified framework through RQ factorizations of deflated matrices at each step. The unified RQ reformulations of these algorithms are easier to understand and implement than the original algorithms.

We also present a new algorithm based on the RQ formulation of the single-input recursive algorithm. We show how this new algorithm differs from other QR algorithms, and establish backward numerical stability of the algorithm through round-off error analysis. In the course of proving backward stability of this algorithm, we prove that any single-input eigenvalue assignment algorithm can be made to be backward stable if it is backward stable for the controller-Hessenberg form.

Finally, we present a detailed round-off error analysis of the single-input recursive algorithm, which is most efficient, and almost trivial to implement, but is numerically suspect. Our forward analysis cannot speak to the stability of the method, but the method is reliable in the sense that we can get an indication, as the method is executed, when the results are suspect, and the indication can be obtained rather cheaply.

The organization of this paper is as follows:

In section 2 we present a unified RQ reformation of the three QR methods.

In section 3 we establish a relationship between these QR methods and the recursive algorithm.

In section 4 we present a new RQ-based algorithm and discuss the differences of this new algorithm with the others.

In section 5 we present the round-off analyses of the proposed algorithm and that of the recursive algorithm.

Finally, in section 6 we present some numerical experiments comparing some of the methods.

2. Hessenberg eigenvalue assignment. The methods to be discussed in this section have the following basic structure: the pair (A, b) is first transformed to a controller-Hessenberg form; the desired feedback is then computed for the reduced problem, and finally the solution to the original problem is retrieved from the solution of the reduced problem. The pair (H, r) is in controller-Hessenberg form if H is an upper Hessenberg matrix and r is a multiple of e_1, the first column of the identity matrix. If r is nonzero and H is unreduced (i.e., $h_{i+1,i} \neq 0, \quad i = 1, 2, \ldots, n-1$), then the system is controllable. The above can be summarized in the following algorithm template.

Algorithm 2.1. A general single-input algorithm.
Input: $A \in \boldsymbol{R}^{n\times n}, b \in \boldsymbol{R}^n$, and $\Omega = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$
Output: $f \in \boldsymbol{R}^n$ such that $\lambda(A - bf^t) = \Omega$

Step 1 Reduce the pair (A, b) to controller-Hessenberg form $(H, \beta e_1) = (PAP^t, Pb)$

Step 2 Compute $k \in \boldsymbol{R}^n$ such that $\lambda(H - \beta e_1 k^t) = \Omega$

Step 3 Compute $f = P^t k$

If in Step 1 it is decided that the system is uncontrollable (i.e., if H is reduced or $\beta = 0$) and if those eigenvalues of A which cannot be moved (called *uncontrollable modes*) do not belong to Ω, then we must stop with failure: Ω is unassignable. If the

uncontrollable modes are contained in Ω, then we go to Step 2 with the controllable part of H and the subset of Ω that remains to be assigned. Since Ω is closed with respect to complex conjugation, then f will be real.

We note here that the (orthogonal) matrix P determined by the reduction must be saved for use in Step 3. Also note that Steps 1 and 3 are individually backward stable operations (see, e.g., [41, pp. 110–160]). We will show in section 5 that a method that is backward stable for Hessenberg eigenvalue assignment problem (Step 2) will be backward stable overall.

Remarks. Several remarks on Algorithm 2.1 are in order.

First, the reduction to controller-Hessenberg form can be achieved in a numerically stable way using a staircase algorithm (see, Boley [7], Paige [31], and Van Dooren and Verhaegen [39]).

Second, Step 1 and Step 3 are the same in all the eigenvalue assignment methods to be discussed in this paper.

The different algorithms differ in the way Step 2 is implemented. We will present below the RQ formulation of several QR-based algorithms, and a recursive algorithm to implement Step 2. We will then present a new algorithm based on the RQ formulation of the recursive algorithm, thus presenting a link between these apparently different algorithms.

Third, in section 5, we will prove that if Step 2 is implemented in a numerically stable way, then the overall algorithm will be numerically stable, thus reproving the numerical stability of several known QR-based algorithms and proving that of the new algorithm. The definition of stability used here is that of Stewart [36].

2.1. The method of Miminis and Paige [27]. The basic idea of the method is to apply the QR algorithm with ultimate shifts to the matrix (with unknown first row) $(H - \beta e_1 k^t)$. If for simplicity we assume that the closed-loop eigenvalues are all real, then the method consists of n deflation steps and a "backward sweep." Each deflation step can be thought of as an RQ factorization of the matrix $(H_i - \beta_i e_1 k_i^t - \lambda_i I)$. However, since k_i is unknown, the process is not quite so straightforward. We first compute Q_i such that $(H_i - \lambda_i I)Q_i^t = R_i$ is upper triangular. Then $U_i = (H_i - \beta_i e_1 k_i^t - \lambda_i I)Q_i^t$ must also be upper triangular, and we want to choose k_i such that U_i is singular. Now since H_i is unreduced, the only way that U_i can be singular is if $U_i e_1 = 0$, that is, if $u_{11}^{(i)} \equiv e_1^t U_i e_1 = 0$. Write $Q_i = \begin{bmatrix} y_i^t \\ \tilde{Q}_i \end{bmatrix}$. Then

$$0 = U_i e_1 = (H_i - \beta_i e_1 k_i^t - \lambda_i I) y_i,$$

or

$$\beta_i k_i^t y_i = e_1^t (H_i - \lambda_i I) y_i = r_{11}^{(i)}. \tag{2.1}$$

This is a key relation in the method, but it does *not* allow us to compute k_i, so we continue. To complete the RQ step we premultiply by Q_i and add back the λ_i to get

$$Q_i (H_i - \beta_i e_1 k_i^t) Q_i^t = \begin{bmatrix} \lambda_i & * \\ 0 & \tilde{Q}_i (H_i - \beta_i e_1 k_i^t) \tilde{Q}_i^t \end{bmatrix}.$$

Now if we define $H_{i+1} = \tilde{Q}_i H_i \tilde{Q}_i^t, \beta_{i+1} = q_{21}^{(i)} \beta_i$, and $k_{i+1} = \tilde{Q}_i k_i$, then the ith deflation step is complete; H_{i+1} is unreduced, β_{i+1} is nonzero, and we can continue with the controllable pair $(H_{i+1}, \beta_{i+1} e_1)$ and the *unknown* feedback vector k_{i+1} of dimension one less than that of k_i.

At the final deflation step we have $H_n - \beta_n e_1 k_n^t \in \boldsymbol{R}^{1\times 1}$, i.e., $k_n = (H_n - \lambda_n I)/\beta_n$ is a real number.

The backward sweep consists of computing $k_{n-1}, k_{n-2}, \ldots, k_1 = k$ using the relations

$$k_{i+1} = \tilde{Q}_i k_i \tag{2.2}$$

and from (2.1)

$$y_i^t k_i = r_{11}^{(i)}/\beta_i. \tag{2.3}$$

Combining these equations we have

$$k_i = Q_i^t \begin{pmatrix} r_{11}^{(i)}/\beta_i \\ k_{i+1} \end{pmatrix}, \quad i = n-1, n-2, \ldots, 1. \tag{2.4}$$

We summarize the preceding discussion as an algorithm.

Algorithm 2.2. The RQ formulation of single-input algorithm of Miminis and Paige.

Input: H, an unreduced $n \times n$ Hessenberg matrix, $\beta \neq 0$, and $\Omega = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$

Output: k such that $\lambda(H - \beta e_1 k^t) = \Omega$

Step 1 Set $H_1 = H$, and $\beta_1 = \beta$
For $i = 1, 2, \ldots, n-1$ do
Compute $(H_i - \lambda_i I)Q_i^t = R_i$, the RQ factorization of $(H_i - \lambda_i I)$
Compute $\tau_i = r_{11}^{(i)}/\beta_i$ and $\beta_{i+1} = q_{21}^{(i)}\beta_i$
Compute H_{i+1}, where $Q_i R_i + \lambda_i I = \begin{bmatrix} * & * \\ 0 & H_{i+1} \end{bmatrix}$
End

Step 2 Compute $k_n = (H_n - \lambda_n)/\beta_n$
For $i = n-1, n-2, \ldots, 1$ do
Compute $k_i = Q_i^t \begin{pmatrix} \tau_i \\ k_{i+1} \end{pmatrix}$
End

Flop count: When implemented with implicit double steps, this algorithm takes about $\frac{5}{6}n^3$ flops. Combined with the $\frac{5}{3}n^3$ flops required for the controller-Hessenberg reduction, the total flop count is about $\frac{5}{2}n^3$.

2.2. The method of Petkov, Christov, and Konstantinov [34]. This method, like the Miminis–Paige method, is based on an ultimately shifted RQ step with immediate deflation. The only real difference between the two methods is *how* the matrices Q_i are computed. In fact, we will show by the end of this section that the Q_i obtained by these methods are essentially the same throughout the entire deflation sequence. We will devote a major portion of this section to an analysis of the RQ factorization (and therefore the deflation step) in the method of Petkov, Christov, and Konstantinov.

If λ is an eigenvalue of the Hessenberg matrix $(H - \beta e_1 k^t)$, then there exists $v \neq 0$ such that

$$(H - \lambda I)v = \beta e_1 k^t v. \tag{2.5}$$

Now partition $(H - \lambda I)$ and v as

$$(H - \lambda I) = \begin{bmatrix} * & * \\ T & c \end{bmatrix} \text{ and } v = \begin{bmatrix} \tilde{v} \\ v_k \end{bmatrix},$$

where $T \in \boldsymbol{R}^{n-1\times n-1}$ is upper triangular. Then from (2.5) we have $[T\ c]v = 0$, or

$$T\tilde{v} = -v_k c.$$

Since H is unreduced, T is nonsingular and v_k is nonzero; so if we fix $v_k \neq 0$, we can compute v by back substitution. We now have an eigenvalue/eigenvector pair (λ, v) of the matrix $(H - \beta e_1 k^t)$, and if we can compute an orthogonal matrix Q such that $Qv = \alpha e_1$ and $Q(H - \lambda I)Q^t$ is a Hessenberg matrix, then

$$0 = (H - \beta e_1 k^t - \lambda I)v = (H - \beta e_1 k^t - \lambda I)\alpha Q^t e_1,$$

or

$$(H - \lambda I)Q^t e_1 = \beta(k^t Q^t e_1)e_1. \tag{2.6}$$

If we now write $Q = \begin{bmatrix} y^t \\ \tilde{Q} \end{bmatrix}$, then (2.6) yields

$$\beta k^t y = e_1^t (H - \lambda I)y.$$

Inserting subscripts and continuing in the fashion of the last section, we see that

$$Q_i(H_i - \beta_i e_1 k_i^t)Q_i^t \quad = \quad \begin{bmatrix} \lambda_i & * \\ 0 & \tilde{Q}_i(H_i - \beta_i e_1 k_i^t)\tilde{Q}_i^t \end{bmatrix}.$$

Define $H_{i+1} = \tilde{Q}_i H_i \tilde{Q}_i^t, \beta_{i+1} = q_{21}^{(i)}\beta_i$, and $k_{i+1} = \tilde{Q}_i k_i$. If Q_i is unreduced, then H_{i+1} is also unreduced, β_i is nonzero, and therefore the pair $(H_{i+1}, \beta_{i+1}e_1)$ is controllable. This is entirely the same situation as in the method of Miminis and Paige, and as such we can use the same backward sweep to recover $k = k_1$.

We have not yet explained how to compute an orthogonal matrix Q such that $Qv = \alpha e_1$ and $Q(H - \lambda I)Q^t$ is a Hessenberg matrix; the following lemma illustrates the construction.

LEMMA 2.1. *Let $Hv = \lambda v$, where H is an unreduced upper Hessenberg matrix. Let the Givens rotations J_k in the k and $(k+1)^{st}$ planes be such that $J_i J_{i+1} \cdots J_{n-1} v = (x_i^t, \alpha_i, 0)^t$ for $i = n-1, n-2, \ldots, 1$, where $x_i \in \boldsymbol{R}^{i-1\times 1}$ and $\alpha_i \in \boldsymbol{R}$. Then*

$$(H - \lambda I)J_{n-1}^t J_{n-2}^t \cdots J_1^t = R \tag{2.7}$$

is upper triangular.

Proof. Define $M_i = (H - \lambda I)J_{n-1}^t J_{n-2}^t \cdots J_i^t$ and suppose that

$$M_i = \begin{bmatrix} k_i & * & * \\ 0 & y_i^t & * \\ 0 & 0 & R_i \end{bmatrix},$$

where $y_i \in R^{2\times 1}, R_i \in R^{n-i\times n-i}$ is upper triangular, and $\begin{bmatrix} k_i & * \\ 0 & y_i^t \end{bmatrix}$ is an unreduced upper Hessenberg matrix of order i. We will show that $\begin{bmatrix} k_{i-1} & * \\ 0 & y_{i-1}^t \end{bmatrix}$ is also an unreduced

upper Hessenberg matrix and that R_{i-1} is upper triangular. Thus, by induction we will have (2.7). Now

$$M_i J_{i-1}^t = \begin{bmatrix} k_i & * & * \\ 0 & \bar{y}_i^t & * \\ 0 & 0 & R_i \end{bmatrix},$$

and since $(H - \lambda I)v = 0$ we must have that

$$0 = M_{i-1}(J_{i-1}J_i \cdots J_{n-1}v) = M_{i-1} \begin{pmatrix} x_{i-1} \\ \alpha_{i-1} \\ 0 \end{pmatrix}.$$

Therefore, $\bar{y}_i^t$ must be of the form $\bar{y}_i^t = (0, y) \in R^{1\times 2}$, with

$$M_{i-1} = \begin{bmatrix} k_{i-1} & * & * \\ 0 & y_{i-1}^t & * \\ 0 & 0 & R_{i-1} \end{bmatrix},$$

R_{i-1} upper triangular, and

$$\begin{bmatrix} k_{i-1} & * \\ 0 & y_{i-1}^t \end{bmatrix}$$

unreduced upper Hessenberg. This completes the induction step; and since M_1 is an unreduced upper Hessenberg matrix, the proof is complete. □

Quite simply, the rotations J_i that are defined by v provide the RQ factorization: $(H - \lambda I)J_{n-1}^t J_{n-2}^t \cdots J_1^t = (H - \lambda I)Q^t = R$.

We summarize the Petkov–Christov–Konstantinov method.

Algorithm 2.3. The RQ formulation of the single–input method of Petkov, Christov, and Konstantinov.

Input: H, an unreduced $n \times n$ Hessenberg matrix, $\beta \neq 0$, and $\Omega = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$

Output: k such that $\lambda(H - \beta e_1 k^t) = \Omega$

Step 1 Set $H_1 = H$, and $\beta_1 = \beta$
For $i = 1, 2, \ldots, n-1$ do
Compute $(H_i - \lambda_i I)Q_i^t = U_i$, the RQ factorization of $(H_i - \lambda_i I)$ by computing v_i such that $(H_i - \lambda_i I)v_i = \gamma e_1$ and then computing the rotations J_i such that $J_1 J_2 \cdots J_{n-i+1} v_i = \pm ||v_i||_2 e_1$, and finally setting $Q_i = J_1 J_2 \cdots J_{n-i+1}$
Compute $\tau_i = e_1^t (H_i - \lambda_i I)Q_i^t e_1/\beta_i$ and $\beta_{i+1} = q_{21}^{(i)}\beta_i$
Compute H_{i+1}, where $Q_i U_i + \lambda_i I = \begin{bmatrix} * & * \\ 0 & H_{i+1} \end{bmatrix}$
End

Step 2 Compute $k_n = (H_n - \lambda_n)/\beta_n$
For $i = n-1, n-2, \ldots, 1$ do
Compute $k_i = Q_i^t \begin{pmatrix} \tau_i \\ k_{i+1} \end{pmatrix}$
End

Flop count: If implemented with care, this algorithm takes about $\frac{5}{3}n^3$ flops. When combined with the $\frac{5}{3}n^3$ flops required for the controller-Hessenberg reduction, the total flop count is about $\frac{10}{3}n^3$.

2.3. The method of Patel and Misra [32]. We have now seen two methods based on an explicit RQ step with immediate deflation. It should come as no surprise that an implicit RQ step is possible, and in order to handle complex pairs of eigenvalues with real arithmetic, an implicit double step is needed. Such a method was proposed by Miminis [28] and Patel and Misra [32]. The method of Patel and Misra is similar to the method of Miminis, but it includes an alternative to the "backward sweep," and is the first published description of the implicit double step in the single-input eigenvalue assignment problem. We will outline an implicit single step here.

First, compute an orthogonal matrix P_i such that $e_n^t(H_i - \lambda_i I)P_i^t = \alpha e_n^t$; then compute another orthogonal matrix U_i such that $U_i P_i H_i P_i^t U_i^t$ is an upper Hessenberg matrix; finally, set $Q_i = U_i P_i$. The matrix U_i "chases the bulge" *up* the subdiagonal of $P_i H_i P_i^t$. We are now in the familiar situation of computing k_i such that

$$Q_i(H_i - \beta_i e_1 k_i^t)Q_i^t = \begin{bmatrix} \lambda_i & * \\ 0 & H_{i+1} - \beta_{i+1} e_1 k_{i+1}^t \end{bmatrix}.$$

If, as before, we set $Q_i = \begin{bmatrix} y_i^t \\ \tilde{Q}_i \end{bmatrix}$, then with $\tau_i \equiv k_i^t y_i$, we must have

$$\tau_i = \frac{\lambda_i - h_{11}^{(i)}}{\beta_i q_{11}^{(i)}} = \frac{h_{21}^{(i)}}{\beta_i q_{21}^{(i)}},$$

with the continuation $H_{i+1} = \tilde{Q}_i H_i \tilde{Q}_i^t, \beta_{i+1} = q_{21}^{(i)}\beta_i$, and $k_{i+1} = \tilde{Q}_i k_i$. After n such steps we expect the usual backward sweep, but it is shown in Patel and Misra [32] that this computation need not be put off that long: the backward sweep

$$k_i = Q_i^t \begin{pmatrix} \tau_i \\ k_{i+1} \end{pmatrix}, \quad i = n-1, n-2, \ldots, 1$$

is equivalent to the "forward update"

$$\hat{Q} = I, \ k = 0, \text{ and}$$

$$k^t = k^t + \tau_i y_i \hat{Q}, \ \hat{Q} = \tilde{Q}_i \hat{Q}, \ i = 1, 2, \ldots, n-1.$$

Algorithm 2.4. The single-input algorithm of Patel and Misra.
Input: H, an unreduced $n \times n$ Hessenberg matrix, $\beta \neq 0$,
and $\Omega = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$
Output: k such that $\lambda(H - \beta e_1 k^t) = \Omega$

Step 1 Set $H_1 = H, \beta_1 = \beta, \hat{Q} = I$, and $k = 0$
For $i = 1, 2, \ldots, n-1$ do
Compute P_i such that $e_n^t(H_i - \lambda_i I)P_i^t = \alpha e_n^t$
Compute U_i such that $U_i P_i H_i P_i^t U_i^t$
is an upper Hessenberg matrix
Set $Q_i = \begin{bmatrix} y_i^t \\ \tilde{Q}_i \end{bmatrix} = U_i P_i$
Compute $\tau = e_1^t(H_i - \lambda_i I)Q_i^t e_1/\beta_i$
Compute $\beta_{i+1} = q_{21}^{(i)}\beta_i$
Compute H_{i+1}, where $Q_i H_i Q_i^t = \begin{bmatrix} * & * \\ 0 & H_{i+1} \end{bmatrix}$
Compute $k = k + \tau\hat{Q}^t y_i^t$
Compute $\hat{Q} = \tilde{Q}_i\hat{Q}$
End

Step 2 Compute $\tau = (H_n - \lambda_n)/\beta_n$
Compute $k = k + \tau\hat{Q}^t$

Flop Count: This method requires about $\frac{5}{6}n^3$ flops for a controller-Hessenberg pair. When combined with the $\frac{5}{3}n^3$ flops required for the controller-Hessenberg reduction, the total flop count is about $\frac{5}{2}n^3$.

2.4. A recursive algorithm [17]. We reproduce below the recursive algorithm of Datta [17], which is apparently different from the three just described, and show how this algorithm produces an explicit formula for the single-input feedback vector.

In the next section, we will present an RQ formulation of this method. This new RQ method will help elucidate the relationship between the other RQ methods and the explicit expression for the feedback vector obtained by the recursive formula.

Algorithm 2.5. A recursive algorithm [17].
Input: H, an unreduced $n \times n$ Hessenberg matrix, $\beta \neq 0$,
and $\Omega = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$
Output: k such that $\lambda(H - \beta e_1 k^t) = \Omega$

Step 1 Set $l_1 = e_n$

Step 2 For $i = 1, 2, \ldots, n-1$ do
Compute $\hat{l}_{i+1} = (H^t - \lambda_i I)l_i$
Compute $l_{i+1} = b_i\hat{l}_{i+1}$, where b_i is chosen
so that $||l_i|| \in (\frac{1}{2}, 1]$, say
End

Step 3 Compute $k = \frac{1}{\beta l_{1n}}(H^t - \lambda_n I)l_n$

Flop count: This method requires only about $\frac{1}{6}n^3$ flops. When combined with the $\frac{5}{3}n^3$ flops required for the controller-Hessenberg reduction, the total flop count is about $\frac{11}{6}n^3$. Given a system in controller-Hessenberg form, this method is more than five times as fast as the method of Miminis and Paige. The assignment of complex pairs of eigenvalues in real arithmetic requires a slight adjustment to the above method but does not alter the operations count.

A closed-form solution of the single-input eigenvalue assignment problem. We now show that this method yields an explicit closed-form solution for the single-input problem. The recursion in Step 2 above yields

$$\gamma l_{i+1} = (H^t - \lambda_1 I)(H^t - \lambda_2 I)\cdots(H^t - \lambda_i I)l_1, \tag{2.8}$$

for some (nonzero) scalar γ. Including Steps 1 and 3, (2.8) becomes

$$\alpha k = (H^t - \lambda_1 I)(H^t - \lambda_2 I)\cdots(H^t - \lambda_n I)e_n, \tag{2.9}$$

where $\alpha = (\beta h_{21} h_{32} \cdots h_{n,n-1})^{-1}$. If $\phi(x) = (x - \lambda_1)(x - \lambda_2)\cdots(x - \lambda_n)$, then this will be written as

$$k^t = \alpha e_n^t \phi(H). \tag{2.10}$$

Since this solution is unique, it represents *the* Hessenberg formula for the single-input eigenvalue assignment problem.

2.5. Of methods not discussed. Varga [40] proposed a method very different from those considered here. It has largely been ignored by numerical linear algebraists because of a reduction of the original system to controller-Schur form ($T = PAP^t$ is block upper triangular with 1×1 or 2×2 diagonal blocks, and $k = Pb$ is a "full" column vector; see Varga [40] for details). It is argued that, besides the extra work involved, the method suffers from the fact that possible ill conditioning of the eigenvalues of the original system introduces unnecessary errors into the computation. These criticisms, while entirely valid from an algorithmic perspective, may be unwarranted from a more macroscopic view. It may be that knowledge of the original spectra (provided by the controller-Schur form and *not* by the controller-Hessenberg form) is necessary for intelligent eigenvalue/partial eigenvalue assignment. In that case the information provided by the Schur decomposition might be used in choosing Ω. If the eigenvalues of the original system were found to be ill conditioned, a Hessenberg method might be preferable, but if not, continuing on with the method of Varga would be more efficient.

There exist many methods for the eigenvalue assignment problem, and we have chosen to discuss only those few with positive numerical attributes (e.g., stability and efficiency). Methods that depend on Jordan or Frobenius forms are both expensive and unstable. Most closed-form solutions for the feedback vector require such forms and hence lead to poor numerical methods. One of the most well-known closed-form solutions is due to Ackermann [1]; while it is often held as an example of how *not* to solve the eigenvalue assignment problem, we will see in the next section that each of the methods discussed in this paper are closely related to that solution.

3. Relationships between the various methods. In this section we will explain the relationships between the methods of Miminis and Paige; Petkov, Christov, and Konstantinov; Patel and Misra; and Datta. We will show that the Miminis–Paige, Petkov–Christov–Konstantinov, and Patel–Misra methods yield the same data at each deflation step, the only difference being the technique used for an RQ factorization. Then we will present an RQ implementation of the recursive method that ties all four of the methods together.

The Miminis–Paige, Petkov–Christov–Konstantinov, and Patel–Misra methods all have an RQ factorization at the heart of the deflation step. With the original method of Miminis and Paige we have the explicit Hessenberg RQ factorization, with

that of Petkov, Christov, and Konstantinov we have a novel "triangular system" Hessenberg RQ factorization, and with the method of Patel and Misra we have the implicit Hessenberg RQ factorization.

These methods all begin with the same data, the pair $(H, \beta e_1)$ and the closed-loop spectrum Ω; furthermore, it is clear that each of the methods generates the $(i+1)$st set of data by applying an RQ iteration step to the ith set of data. Thus, given the matrix H_i, the implicit-Q theorem (or the uniqueness of the RQ factorization) guarantees that whichever method we choose, the unreduced Hessenberg matrix H_{i+1} is essentially (that is, up to a diagonal scaling of ± 1) the same. One might question the uniqueness of the RQ factorization (or equivalently, the implicit RQ step) if λ_i is an eigenvalue of H_i. Indeed, in this case it is not unique, for while Q_i *is completely determined*, the first row of R_i is underdetermined. But if we now note that the deflation step is taken immediately in each of the methods, it is clear that the first row of R_i *plays no part in the computation*. We have proven the following lemma.

LEMMA 3.1. *In exact arithmetic, the methods of Miminis and Paige, Petkov, Christov, and Konstantinov, and Patel and Misra all generate the same data H_i, and Q_i at each deflation step, up to a sign scaling, for $i = 1, 2, \ldots, n$.*

The differences between these methods, *at each* step, depend only on finite precision. The discussion above allows us to give a generic formulation of all of the QR-based single-input algorithms as follows.

Algorithm 3.1. Generic RQ-based single-input algorithm.

Input: H, an unreduced $n \times n$ Hessenberg matrix, $\beta \neq 0$, and $\Omega = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$

Output: k such that $\lambda(H - \beta e_1 k^t) = \Omega$

Step 1 Set $H_1 = H$ and $\beta_1 = \beta$
For $i = 1, 2, \ldots, n-1$ do
Compute Q_i from a shifted RQ step with H_i and λ_i:
$\tilde{H}_i = Q_i H_i Q_i^t$
Compute $\tau_i = \frac{\tilde{h}_{21}^{(i)}}{\beta_i q_{21}^{(i)}} = \frac{\lambda_i - \tilde{h}_{11}^{(i)}}{\beta_i q_{11}^{(i)}}$
Compute $\beta_{i+1} = \beta_i q_{21}^{(i)}$
Compute H_{i+1}, where $\tilde{H}_i = \begin{bmatrix} * & * \\ 0 & H_{i+1} \end{bmatrix}$
End

Step 2 Compute $\tau_n = \frac{\lambda_n - H_n}{\beta_n}$

Step 3 Compute $k^t = (\tau_1, \tau_2, \ldots, \tau_n) Q_{n-1} Q_{n-2} \cdots Q_1$

The manner of computing the RQ steps in this generic method is not specified; an explicitly shifted RQ step with Givens rotations yields the original method of Miminis–Paige, explicitly computing the RQ factors using a closed-loop eigenvector gives the method of Petkov–Christov–Konstantinov, and an implicit RQ step corresponds to the methods of Patel–Misra and Miminis–Paige. We also note that τ_i can be computed using either of the quantities given above, or if R_i is available, as $\tau_i = r_{11}^{(i)}/\beta_i$.

4. A new RQ-based method. We now present a new RQ implementation of the recursive algorithm of Datta that will make explicit the connections between all of these methods and the explicit formula (2.10).

While this method was discovered and proved in the context of the matrix equation

$$H^tL - LB = ce_n^t,$$

we can show its relationship with the often-maligned formula of Ackermann. Ackermann [1] showed that if

$$\phi(x) = (x - \lambda_1)(x - \lambda_2)\cdots(x - \lambda_n),$$

then the unique solution to the eigenvalue assignment problem for the controllable pair (A, b) is

$$f^t = e_n^t C^{-1}\phi(A), \tag{4.1}$$

where

$$C \equiv \left[\begin{array}{cccc} b, & Ab, & \ldots, & A^{n-1}b \end{array}\right].$$

If $(H, \beta e_1) = (PAP^t, Pb)$ is the controller-Hessenberg form of (A, b), then from (4.1)

$$\begin{aligned} f^tP^t &= e_n^tC^{-1}\phi(A)P^t \\ &= e_n^tC^{-1}P^t\phi(H) \\ &= e_n^t(PC)^{-1}\phi(H), \end{aligned}$$

where PC is an upper triangular matrix. If α^{-1} is the (n, n) element of PC, then $e_n^t(PC)^{-1} = \alpha e_n^t$, and we see that the formula, $k^t = \alpha e_n^t\phi(H)$ is a Hessenberg case of Ackermann's formula (in fact $\alpha^{-1} = \beta\prod_{i=1}^{n-1} h_{i+1,i}$).

The recursive algorithm is an extremely efficient way to solve the Hessenberg single-input problem, but as shown in [2], backward stability cannot be guaranteed. Having been aware of possible instabilities in the recursive formulation, Datta [20] suggested that this method could be implemented using QR iterations as follows:
Set $H_1 = H$
For $i = 1, 2, \ldots, n$ compute the QR step
$\quad Q_iR_i \;:=\; H_i - \lambda_iI$
$\quad H_{i+1} \;:=\; R_iQ_i + \lambda_iI$
Then it can be shown [36, p. 353] that

$$\phi(H) = Q_1Q_2\cdots Q_nR_nR_{n-1}\cdots R_1,$$

and setting $Q = Q_1Q_2\cdots Q_n$ and $R = R_nR_{n-1}\cdots R_1$, formula (2.10) becomes

$$k^t = \alpha e_n^tQR.$$

The difficulty of implementing this strategy is that the R_i need to be accumulated; this is both expensive and unstable.

We now show how the method can be made computationally efficient by using RQ factorizations instead of the QR factorizations:
Set $H_1 = H$
For $i = 1, 2, \ldots, n$ compute the RQ step
$\quad R_iQ_i \;:=\; H_i - \lambda_iI$
$\quad H_{i+1} \;:=\; Q_iR_i + \lambda_iI$
This time

$$\phi(H) = R_1R_2\cdots R_nQ_nQ_{n-1}\cdots Q_1, \tag{4.2}$$

and by setting $Q = Q_n Q_{n-1} \cdots Q_1$ and $R = R_1 R_2 \cdots R_n$, we have

$$k^t = \alpha e_n^t RQ \quad = \quad \alpha \rho e_n^t Q,$$

where $\rho = \prod_{i=1}^{n} r_{nn}^{(i)}$. This is a much nicer situation! Furthermore, we will now show that it is possible to "deflate" the problem at each RQ step.

Write $Q_i, i = 1, 2, \ldots, n$ as a product of Givens rotations $Q_i = J_1^{(i)} J_2^{(i)} \cdots J_{n-1}^{(i)}$, where $J_k^{(i)}$ is a rotation in the k and $k+1$ planes. Then

$$\begin{aligned} e_n^t Q &= e_n^t J_1^{(n)} J_2^{(n)} \cdots J_{n-1}^{(n)} Q_{n-1} Q_{n-2} \cdots Q_1 \\ &= e_n^t J_{n-1}^{(n)} Q_{n-1} Q_{n-2} \cdots Q_1 \\ &= e_n^t J_{n-1}^{(n)} J_{n-2}^{(n-1)} J_{n-1}^{(n-1)} Q_{n-2} Q_{n-3} \cdots Q_1 \\ &\vdots \\ &= e_n^t (J_{n-1}^{(n)})(J_{n-2}^{(n-1)} J_{n-1}^{(n-1)}) \cdots (J_1^{(2)} J_2^{(2)} \cdots J_{n-1}^{(2)}) Q_1, \end{aligned} \tag{4.3}$$

or $k^t = \alpha \rho e_n^t \hat{Q}_n \hat{Q}_{n-1} \cdots \hat{Q}_1$, where $\hat{Q}_k \equiv J_{k-1}^{(k)} J_k^{(k)} \cdots J_{n-1}^{(k)}$. Algorithmically we have the following.

Algorithm 4.1. A proposed single-input algorithm.

Input: H, an unreduced $n \times n$ Hessenberg matrix, $\beta \neq 0$, and $\Omega = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$

Output: k such that $\lambda(H - \beta e_1 k^t) = \Omega$

Step 1 Compute $R_1 Q_1 := H - \lambda_1 I$, the RQ factorization of $H - \lambda_1 I$
Compute $H_2 = Q_1 R_1 + \lambda_1 I$
Set $Q = Q_1$ and $\rho = r_{nn}^{(1)}$

Step 2 For $i = 2, 3, \ldots, n-1$
Compute $R_i Q_i := H - \lambda_i I$
Compute H_{i+1}, where $Q_i R_i + \lambda_i I = \begin{bmatrix} * & * \\ 0 & H_{i+1} \end{bmatrix}$
Update $Q := \begin{bmatrix} I & \\ & Q_k \end{bmatrix} Q$
Update $\rho := \rho r_{n+2-i,n+2-i}^{(i)}$ ($r_{n+2-i,n+2-i}^{(i)}$ is the last element of R_i)
End

Step 3 Update $\rho := \rho(H_n - \lambda_n)$
Compute $k^t = \alpha \rho e_n^t Q$

Remark. The RQ factorizations in this method can be implemented implicitly with a double step, but as with the previous methods this has been omitted for the sake of clarity. Note also that there are no divisions in this method until α is computed in the last step, i.e., the computation of

$$\tau_i = \frac{\tilde{h}_{21}^{(i)}}{\beta_i q_{21}^{(i)}} = \frac{\lambda_i - \tilde{h}_{11}^{(i)}}{\beta_i q_{11}^{(i)}} = \frac{r_{11}^{(i)}}{\beta_i}$$

which appeared in the other methods does not appear here.

Flop count: Implemented using implicit double steps with Q kept in factored form, this method requires about $\frac{5}{6}n^3$ flops. When combined with the $\frac{5}{3}n^3$ flops required for the controller-Hessenberg reduction, the total flop count is about $\frac{5}{2}n^3$, the same as the Miminis–Paige and Patel–Misra methods.

4.1. A relationship between the proposed method and other RQ methods. We promised that this method would shed some light on the relationship between the other RQ methods and the closed-form solution (2.10). While the connection between this RQ method and the closed-form solution is clear, we have yet to close the final link. There are two differences between this method and the generic RQ method: (i) deflation does not commence after the first iteration here as in the generic method, and (ii) the scalar $\alpha\rho$ in this method takes the place of the vector $x^t = (\tau_1, \tau_2, \ldots, \tau_n)$.

When viewed from the perspective of the generic method, these two distinctions are the result of transforming the vector x^t into the vector $\alpha\rho e_n^t$, one step at a time.

To see how this works let us consider an explicit RQ factorization of H_i, an unreduced Hessenberg matrix of order k, say. In the generic method the RQ factorization effectively stops when the matrix $H_iQ_i^t$ is of the form

$$H_iQ_i^t = \begin{bmatrix} A & * \\ 0 & T \end{bmatrix}, \tag{4.4}$$

where A is 2×2 and T is upper triangular. In the proposed RQ method above, we are one deflation step behind, so that one more rotation is needed to put A into triangular form. This rotation V_i will be such that $H_iQ_i^tV_i^t$ is upper triangular, but it also rolls $[\tau_i, \tau_{i+1}]$ into $[0, \gamma]$, for

$$\begin{aligned} [a_{21}, a_{22}] &= [r_{11}^{(i)}\tfrac{\beta_{i+1}}{\beta_i}, r_{11}^{(i+1)}] \\ &= \beta_{i+1}[\tfrac{r_{11}^{(i)}}{\beta_i}, \tfrac{r_{11}^{(i+1)}}{\beta_{i+1}}] \\ &= \beta_{i+1}[\tau_i, \tau_{i+1}]. \end{aligned} \tag{4.5}$$

We have proven the following theorem.

THEOREM 4.1. *The generic RQ method generates the orthogonal matrices Q_i that satisfy*

$$k^t = (\tau_1, \tau_2, \ldots, \tau_n)Q_nQ_{n-1}\cdots Q_1,$$

while the proposed method generates the orthogonal matrices $P_i = Q_iV_i$ such that

$$k^t = \alpha\rho e_n^tP_nP_{n-1}\cdots P_1,$$

where

$$(\tau_1, \tau_2, \ldots, \tau_n)V_1V_2\cdots V_{n-1} = \alpha\rho e_n^t$$

and V_i is a rotation in the planes i and $i+1$.

5. Error analysis. A systematic round-off error analysis of most of the existing and currently used algorithms in control theory is lacking. As far as algorithms for the eigenvalue assignment problem are concerned, round-off error analyses of only the methods of Miminis and Paige and Petkov, Christov, and Konstantinov have been presented (Cox and Moss [13, 14], Miminis and Paige [29]).

In this section we give a detailed round-off error analysis of our proposed single-input algorithm (Algorithm 4.1) described in section 4, and prove that it is backward stable. In the course of this proof we show that any algorithm for the eigenvalue assignment problem is backward stable if it is backward stable for the corresponding Hessenberg problem. We then give a round-off error analysis of the recursive algorithm. Our analysis shows that the latter, while it may not be backward stable, is reliable in the sense that we can detect precisely when the results are suspect.

5.1. An error analysis of the proposed single-input method. The backward error analysis of eigenvalue assignment methods has turned out to be a nontrivial task. For example, the RQ-based methods of section 2 are straightforward adaptations of the Hessenberg QR iteration, and while a backward error analysis for the QR iteration is quite simple, that for the eigenvalue assignment methods is not (see, e.g., Cox and Moss [13, 14]). The major difference is that backward error analysis for the QR iteration in the eigenvalue problem is naturally focused on showing that the next iterate is (exactly) similar to a matrix that is close to the current iterate, while for eigenvalue assignment similarity cannot be used in such a direct, sequential fashion. In order to simplify the analysis, we show that backward stability is achieved if one can solve the Hessenberg single-input eigenvalue assignment problem in a backward stable manner. First we prove that Algorithm 4.1 is backward stable.

THEOREM 5.1. *The RQ-based single-input eigenvalue assignment (Algorithm* 4.1*) is backward stable, i.e., it computes a feedback* k *such that*

$$\lambda(H + \delta H - (\beta + \delta\beta)e_1(k + \delta k)^t) = \Omega,$$

where δH, $\delta\beta$, *and* δk *are small.*

Proof. Let $\bar{H}_1 = H_1 = H$, and let $\bar{H}_i$ be the computed iterate at the ith QR step. Let $\bar{Q}_i$ be the computed transformation at each step. Then we have from basic error analysis (see, e.g., [41, pp. 110–160]) that there exists an orthogonal matrix Q_1 such that

$$\bar{H}_2 = Q_1(H_1 + \delta H_1)Q_1^t, \quad \bar{Q}_1 + E_1 = Q_1,$$

where $||\delta H_1||_F \leq f(n)\boldsymbol{u}\max\{||H_1||_F, |\lambda_i|\}$, and $||E_1||_F \leq g(n)\boldsymbol{u}$, where g and f are modest functions of n, practically behaving like cn, with c a constant of order unity. Note that with an implicit double RQ step, the upper bound on $||\delta H_1||_F$ is independent of λ_i. Now iterating on these results leads to

$$\bar{H}_n = P_{n-1}(H_1 + \delta H)P_{n-1}^t, \quad \bar{P}_{n-1} + E = P_{n-1}, \tag{5.1}$$

where $P_{n-1} = Q_{n-1}Q_{n-2}\cdots Q_1$, $\bar{P}_{n-1} = fl(\bar{Q}_{n-1}\bar{Q}_{n-2}\cdots\bar{Q}_1)$. Here, $||\delta H||_F \leq \boldsymbol{u}nf(n)\ \max\{||H||_F, |\lambda_k|, k = 1, 2, \ldots, n\}$ and $||E||_F \leq ng(n)\boldsymbol{u}$. These bounds are pessimistic now because of the maximum over $|\lambda_k|$ and because we have not considered the fact that H_i and Q_i are actually $(n-i+1)\times(n-i+1)$, not $n\times n$. Now with $P = P_n$, the feedback k is computed as $k^t = \gamma e_n^t P$, so up to the scalar γ, we are done. We have shown that $e_n^t(\bar{P} + E)$ is exact for a matrix $H + \delta H$, where $||e_n^t E|| \leq ng(n)\boldsymbol{u}$ and $||\delta H|| \leq \boldsymbol{u}nf(n)\ \max\{||H||_F, |\lambda_k|\}$. We now show that the scalar γ can be computed in a backward stable fashion, thereby completing the proof.

Remember that $\gamma = \alpha\rho$, where $\rho = \prod_{i=1}^n r_i$, $\alpha = (\prod_{i=1}^n \beta_i)^{-1}$, $\beta_i = h_{i,i-1}$, $\beta_0 = \beta$, and $r_i = r_{nn}^{(i)}$ is the (n, n) entry of R_i. Now let $\bar{r}_i$ be the computed value of r_i, where R_i is exact for the computed matrix $\bar{H}_i$. Then we have

$$\bar{r}_i = \pm\sqrt{\{h_{n,n-1}^2(1 + \epsilon_1) + [(h_{nn} - \lambda_i)(1 + \epsilon_2)]^2(1 + \epsilon_3)\}(1 + \epsilon_4)}(1 + \epsilon_5),$$

where $|\epsilon_j| < \boldsymbol{u}$. Write this as $\bar{r}_i = r_i(1+\delta_i)$, and using $\sqrt{1+\epsilon} = 1 + \epsilon/2 + O(\epsilon^2)$, we have $|\delta| \leq 3\boldsymbol{u} + O(\epsilon^2)$. Now $\bar{\gamma} = fl(\bar{\rho}\bar{\alpha})$ so that

$$\bar{\gamma} = \prod_{i=1}^{n} \frac{r_i(1+\delta_i)(1+\tau_i)}{\beta_i(1+\epsilon_i)},$$

where $|\tau_i|, |\epsilon_i| \leq \boldsymbol{u}$, $i = 1, 2, \ldots, n$. Therefore,

$$\gamma - \bar{\gamma} = \gamma \prod_{i=1}^{n} \frac{(1+\delta_i)(1+\tau_i)}{(1+\epsilon_i)},$$

and if we assume that $n\boldsymbol{u} < 0.1$, then conservatively

$$|\gamma - \bar{\gamma}| \leq 5n\boldsymbol{u}|\gamma|.$$

If we write $\gamma = \gamma(\beta, H)$, then our result reads

$$\bar{\gamma} = \gamma(\beta + \hat{\delta\beta}, H + \delta H),$$

where δH is the same as in (5.1), and $|\hat{\delta\beta}| \leq 5n\boldsymbol{u}$. Finally, the error from the scalar-vector operation $k^t = \gamma e_n^t P$ can be thrown back into β, yielding a computed feedback $\bar{k}$ such that

$$\bar{k} = \bar{\gamma} e_n^t E + k(\beta + \delta\beta, H + \delta H),$$

with $|\delta\beta| \leq 5n\boldsymbol{u}$. □

Remark. The popular definition for backward stability is not used here for a very simple reason. Consider proving that the computation of a Householder reflection is backward stable. One must show that the computed matrix is exact for a problem close to the original. This is impossible, even for the $n = 2$ case, for the computed matrix is almost always *not an orthogonal matrix.* This difficulty is removed by adopting the more general definition of Stewart [36, p. 76], which requires that the computed solution be "near the exact solution of a slightly perturbed problem." Datta [21, p. 87] has called such stability *mild stability.* In the above proof, the quantity $\bar{\gamma} e_n^t E$ is the difference between the computed solution and the exact solution for the perturbed problem, and with $k \equiv k(\beta + \delta\beta, H + \delta H)$ we have (pessimistically) $\|\bar{k} - k\|/\|k\| \leq cn^2\boldsymbol{u}$, where c is a constant of order unity.

THEOREM 5.2. *The following three-step procedure for solving the controllable single-input eigenvalue assignment problem for* (A, b, Ω) *is backward stable if Step* 2 *is backward stable in the mild sense.*

Step 1 *Using the method of Householder, reduce the pair* (A, b) *to the controller-Hessenberg form* $(H, r) = (QAQ^t, Qb)$.

Step 2 *Compute the solution* k *to the eigenvalue assignment problem for* (H, r, Ω).

Step 3 *Compute* $f = Q^t k$.

Proof. Let $\bar{H}$, $\bar{r}$, and $\bar{Q}$ be the computed versions of H, r, and Q, respectively. There exists an orthogonal matrix $\hat{Q}$ such that

$$\hat{Q} - \bar{Q} = E_Q, \ \hat{Q}A\hat{Q}^t = \bar{H} + \hat{E}_H, \text{ and } \hat{Q}b = \bar{r} + \hat{\epsilon}_r, \tag{5.2}$$

where $\|E_Q\| \leq \boldsymbol{u} y_0(n)$, $\|\hat{E}_H\| \leq 2cn\boldsymbol{u}\|A\|$, $\hat{\epsilon}_r \leq cn\boldsymbol{u}\|b\|$, and y_0 behaves, for all practical purposes, like $cn^{3/2}$, with c of order 10 [41, pp. 160–161].

Let $\bar{k}$ be the computed solution to the reduced problem $(\bar{H}, \bar{r}, \Omega)$. Denote by $\bar{f}$ the computed solution to the original problem; then

$$\bar{f} \equiv \mathrm{fl}(\bar{Q}^t \bar{k}) = \bar{Q}^t \bar{k} + \bar{\epsilon}_f, \tag{5.3}$$

where $\|\bar{\epsilon}_f\| \leq n\boldsymbol{u}\|\bar{k}\|$. Substituting (5.2) into (5.3), we have

$$\bar{f} = \hat{Q}^t(\bar{k} + \hat{\epsilon}_k). \tag{5.4}$$

Here $\hat{\epsilon}_k = \bar{\epsilon}_f - E_Q^t \bar{k}$, and so $\|\hat{\epsilon}_k\| \leq \boldsymbol{u} n y_1(n)\|\bar{k}\|,\ y_1(n) = y_0(n) + n$.

Now by hypothesis there exist $(\tilde{H}, \tilde{r})$ close to $(\bar{H}, \bar{r})$, and $\tilde{k}$ close to $\bar{k}$, such that $\tilde{k}$ is the exact solution to the eigenvalue assignment problem with input $(\tilde{H}, \tilde{r}, \Omega)$. Write $\bar{H} = \tilde{H} + \tilde{E}_H,\ \bar{r} = \tilde{r} + \tilde{\epsilon}_r$, and $\bar{k} = \tilde{k} + \tilde{\epsilon}_k$, where

$$\|\tilde{E}_H\| \leq \boldsymbol{u} z_H \|A\|,\ \|\tilde{\epsilon}_r\| \leq \boldsymbol{u} z_r \|b\|, \text{ and } \|\tilde{\epsilon}_k\| \leq \boldsymbol{u} z_k \|\bar{k}\|.$$

Finally, taking $\delta A \equiv \hat{Q}^t(\tilde{E}_H + \hat{E}_H)\hat{Q},\ \delta b \equiv -\hat{Q}^t(\tilde{\epsilon}_r + \hat{\epsilon}_r)$, and $\delta \bar{f} = -\hat{Q}^t(\tilde{\epsilon}_k + \hat{\epsilon}_k)$, we have

$$A + \delta A - (b + \delta b)(\bar{f} + \delta \bar{f})^t = \hat{Q}^t(\tilde{H} - \tilde{b}\tilde{k}^t)\hat{Q}, \tag{5.5}$$

with

$$\|\delta A\| \leq \boldsymbol{u}\|A\|(z_H + 2cn), \|\delta b\| \leq \boldsymbol{u}\|b\|(z_r + cn), \text{ and } \|\delta f\| \leq \boldsymbol{u}\|\bar{k}\|(z_k + y_1(n)). \qquad \square$$

Remark. Since our proposed Hessenberg algorithm is backward stable in the more general sense, the above theorem guarantees that our method is backward stable (in the standard sense, if one prefers) for the original problem.

5.2. An error analysis of the recursive single-input method. Recall that this method computes a matrix L and a vector k such that $HL - L\Lambda = ke_n^t$. A careful look at the iteration reveals that the forward error has a special form. Define the polynomials $\phi_{j,k}$ for $j \leq k$ by

$$\phi_{j,k}(x) = (x - \lambda_j)(x - \lambda_{j+1}) \cdots (x - \lambda_k).$$

THEOREM 5.3. *Let $\bar{\alpha}\bar{f}$ be the* computed *solution to the single-input eigenvalue assignment problem for $(H, \beta e_1, \Omega)$ using the recursive method. If αf is the exact solution, then*

$$\bar{\alpha}\bar{f} - \alpha f = \sum_{j=1}^{n} \phi_{j,n}(H)\epsilon_j. \tag{5.6}$$

Proof. Let $\bar{l}_i$ be the computed value of the ith column of L. Define ϵ_i by $\bar{l}_{i+1} = (H - \lambda_i I)\bar{l}_i + \epsilon_i$. Since $\bar{l}_1 = l_1$, we must have that $\bar{l}_2 = l_2 + \epsilon_1$; suppose $\bar{l}_i = l_i + \sum_{j=1}^{i-1} \phi_{j,i-1}(H)\epsilon_j$. Then

$$\begin{aligned} \bar{l}_{i+1} &= (H - \lambda_i I)\bar{l}_i + \epsilon_i \\ &= (H - \lambda_i)(l_i + \textstyle\sum_{j=1}^{i-1} \epsilon_j \phi_{j,i-1}(H)) + \epsilon_i \\ &= l_{i+1} + \textstyle\sum_{j=1}^{i} \phi_{j,i}(H)\epsilon_j. \end{aligned}$$

Now $\bar{\alpha}\bar{f} = \bar{l}_{n+1}$, and therefore

$$\bar{\alpha}\bar{f} = l_{n+1} + \sum_{j=1}^{n} \phi_{j,n}(H)\epsilon_j,$$

or

$$\bar{\alpha}\bar{f} - \alpha f = \sum_{j=1}^{n} \phi_{j,n}(H)\epsilon_j. \qquad \square$$

The ϵ_j can easily be bounded; for example, if a machine base scaling is used to normalize $\bar{l}_j$, then it is simple to show that

$$||\epsilon_j||_F \leq \beta_m n \boldsymbol{u} ||H - \lambda_j||_F,$$

where β_m is the base. Unfortunately, not much can be said about backward stability from a result like this. It is not a necessarily bad result either, for the closed-form expression for the single-input feedback is $\alpha e_n^t \phi_{1,n}(H)$.

It is possible to shed some light on the stability of this method by looking at the ϵ_j in a different way.

THEOREM 5.4. *Let $E = [\epsilon_1, \epsilon_2, \ldots, \epsilon_n]$ and let $\bar{L} = \left[\bar{l}_1, \bar{l}_2 \ldots, \bar{l}_n\right]$. Then $\bar{\alpha}\bar{f}$ solves (exactly) the single-input eigenvalue assignment program for the perturbed system $(H - E\bar{L}^{-1}, \beta e_1, \Omega)$, where the ϵ_i are the same as in Theorem* 5.3.

Proof. Notice that, as defined, $\bar{L}$ satisfies the Sylvester equation

$$H\bar{L} - \bar{L}\Lambda = E + \bar{\alpha}\bar{f}e_n^t,$$

where $\Lambda = \text{diag}(\lambda_i)$. Since $\bar{L}$ is nonsingular by construction, we can solve the perturbed equation

$$(H + \Delta H)\bar{L} - \bar{L}\Lambda = E + \bar{\alpha}\bar{f}e_n^t \tag{5.7}$$

for ΔH. This yields $-\Delta H = E\bar{L}^{-1}$, and by satisfying (5.7), $\bar{\alpha}\bar{f}$ solves the eigenvalue assignment program for $(H + \Delta H, \beta e_1, \Omega)$. $\square$

5.3. Remarks on numerical stability and reliability. From the above result we cannot say that the method is backward stable. We have simply provided an upper bound on the size of the ball around the initial data, inside which there exist $(H + \Delta H, \beta + \delta\beta)$ for which the computed solution is exact. If $||\Delta H||$ could be bounded above by a small quantity that was relatively independent of the initial data, then the method would be backward stable. But Theorem 5.4 does allow one to say precisely when the results from the method are suspect. It is clear that $||E||$ is always small if the iterates are normalized every few steps, so that all of the backward error information is contained in $\bar{L}^{-1}$. Since $\bar{L}$ is triangular, it is possible to estimate $||\bar{L}^{-1}||$ rather cheaply, even as the iteration proceeds.

The matrix L yields a bit more information about the eigenvalue assignment problem. If the closed-loop eigenvalue problem is poorly conditioned, then we cannot expect the closed-loop eigenvalues to be correct (or even well defined), even when the feedback f is computed to very high accuracy. Now we know from construction that the method yields matrices L and Λ such that

$$H - \beta e_1 f^t = L\Lambda L^{-1},$$

where Λ is bidiagonal. It is easy to show that if the closed-loop eigenvalues are distinct, then Λ is diagonalized by the matrix $X = [x_{ij}]$, where

$$x_{ij} = \begin{cases} 1, & i = j, \\ \prod_{k=1}^{j-1}(\lambda_k - \lambda_j)^{-1}, & j > i, \\ 0, & j < i. \end{cases}$$

Therefore, the closed-loop matrix is diagonalized by the matrix $P = L^{-1}X$ which is conveniently factorized into triangular factors, with X a unit upper triangular matrix. The inverse of P is given by $P^{-1} = X^{-1}L$, where $X^{-1} = [y_{ij}]$, and

$$y_{ij} = \begin{cases} 1, & i = j, \\ \prod_{k=i+1}^{j} (\lambda_j - \lambda_k)^{-1}, & j > i, \\ 0, & j < i. \end{cases}$$

This leads us to an upper bound on the eigencondition of the closed-loop matrix

$$\|P\|\|P^{-1}\| = \|L^{-1}X\|\|X^{-1}L\| \leq \|X\|\|X^{-1}\|\|L\|\|L^{-1}\|.$$

The triangular factors facilitate an $O(n^2)$ LINPACK-like condition estimator of $P = L^{-1}X$. We cannot say that whenever L is illconditioned, the closed-loop eigenvalues are ill-conditioned, for L is simply a factor of P, but computational experience has shown that it is a good indicator.

Numerical experiments. We include here several computational experiments that compare the accuracy of the proposed method (RQ) with that of Miminis and Paige (M&P) and Datta. The M&P method was chosen as representative of the QR-based methods primarily because of the MATLAB script SEVAS, written by Miminis, and available to the public [30]. All computations were done on a SUN Sparcstation LX. MATLAB, version 4.2C [26], was used to compute the feedback vector using the m-files SEVAS.m for the M&P method, SIPPD.m for Datta's method, and SIPPRQ.m for the proposed method (SIPPD.m and SIPPRQ.m available from Arnold). MATLAB computations are double precision with a machine epsilon of $\mu = 2^{-52}$. In all tests an "exact" feedback was computed using the method of Datta, coded in Bailey's multiprecision FORTRAN [6] with a 500 decimal digit floating point representation. Datta's method was chosen for its efficiency and ease of implementation. In all of the experiments, the initial data is in controller-Hessenberg form. The computation of an exact solution allows one to avoid the eigenvalue computation (and the associated errors) necessary in the common practice of measuring error by computing the eigenvalues of H, removing its first row, and then assigning the original eigenvalues to the perturbed matrix.

For a backward stable method, one expects the size of the error in the computed solution to be roughly equal to the product of the machine epsilon and the condition number of the problem. We have included in these tests the computation of a relative condition estimator (the estimator ν_ϕ, given in [2], requires about $\frac{1}{15}$ the work of either of the methods being compared). We would like to emphasize two points here: first, we are measuring the error in the computed feedback, not in the closed-loop eigenvalues; and second, this condition estimator is neither a lower nor upper bound on the true condition number, which, while computable, requires at least $O(n^4)$ flops for the general case.

For all of these experiments the MATLAB code that generates the test data, and the seeds for the random number generator, are available from Arnold.

In the first experiment, a random matrix with elements uniformly distributed in $[-1, 1]$ was generated using MATLAB's RAND function. This matrix was then reduced to Hessenberg form and its elements rounded to 15 binary digits, resulting in the system matrix H. Next, a unit random vector r was generated and the eigenvalues

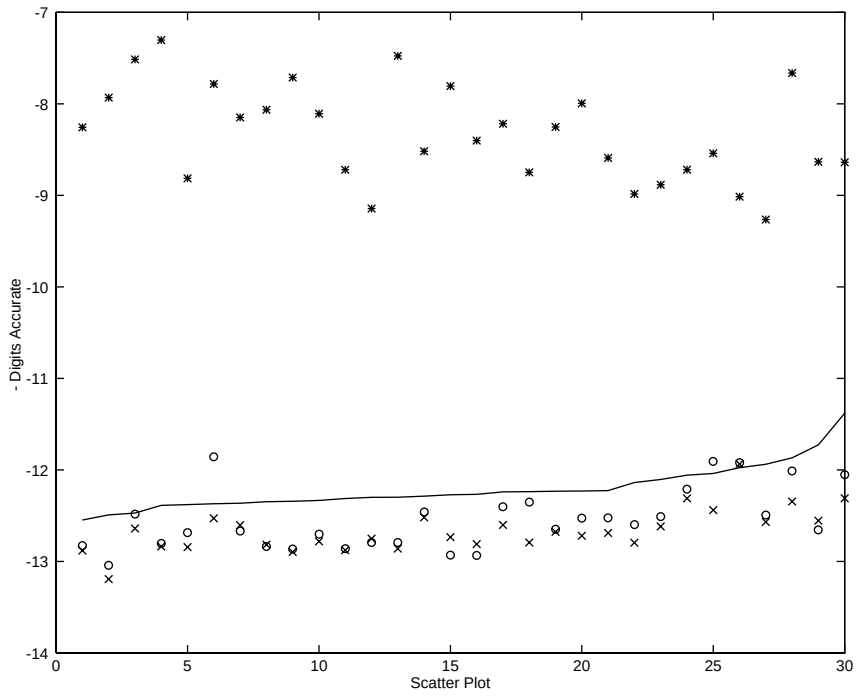


FIG. 1. *Scatter plot for* 30 *problems of size* 100. *$log_{10}(e_c)$ is (the negative of) the number of correct decimal digits in the computed feedback. The predicted error is given by the "continuous" curve, Datta's method is represented by "*", MP by "o", and the RQ method by "x".*

of the matrix $H - e_1 r^t$ computed. These eigenvalues, rounded to 15 binary digits, become the desired closed-loop poles. For a relatively well-conditioned eigenvalue assignment problem, we expect the exact feedback to have norm near unity.

Thirty such runs were performed on matrices of size $n = 100$. The results are decribed in Figure 1 and Table 1. Figure 1 is a scatter plot showing $-\log_{10}(e_c)$, where $e_c = \|f - f_c\|/\|f\|$, f is the exact feedback, and f_c is the feedback computed by one of the methods being compared. The x-axis serves only to label the data points; each integer k, from 1 to 30, represents a data point, and each data point consists of four quantities, namely, the predicted error and the error for each of the three methods being compared. The y-axis in the figure represents the (negative of) the number of correct digits in the computation, thus a smaller (closer to $-\infty$) y-component represents a smaller error. In order to make the plot easier to read the data is sorted by the predicted error $\mu\nu_\phi$ and the predicted error is plotted as a continuous curve by linear interpolation. Note that even for problems of size $n = 100$ (considered large for single-input eigenvalue assignment), the feedback vector is computed to high relative accuracy by the backward stable methods. This observation supports the argument that the generically dismal behavior of eigenvalue assignment for large n is not caused by innacurate feedback but is primarily attributable to the conditioning of the closed-loop eigenvalues relative to the size of the feedback vector.

Table 1 provides some statistics associated with the data shown in Figure 1. The quantity "digits accurate" is simply $-\log_{10}(e_c)$, which is approximately the number

TABLE 1
Summary statistics of relative errors for 30 *randomly generated systems of order* 100.

	Accurate Digits			
Method	*Average*	*Minimum*	*Backward average*	*Backward minimum*
M&P	12.5	11.9	16.0	15.1
RQ	12.7	11.9	16.1	15.6
Datta	8.33	7.30	11.8	10.6

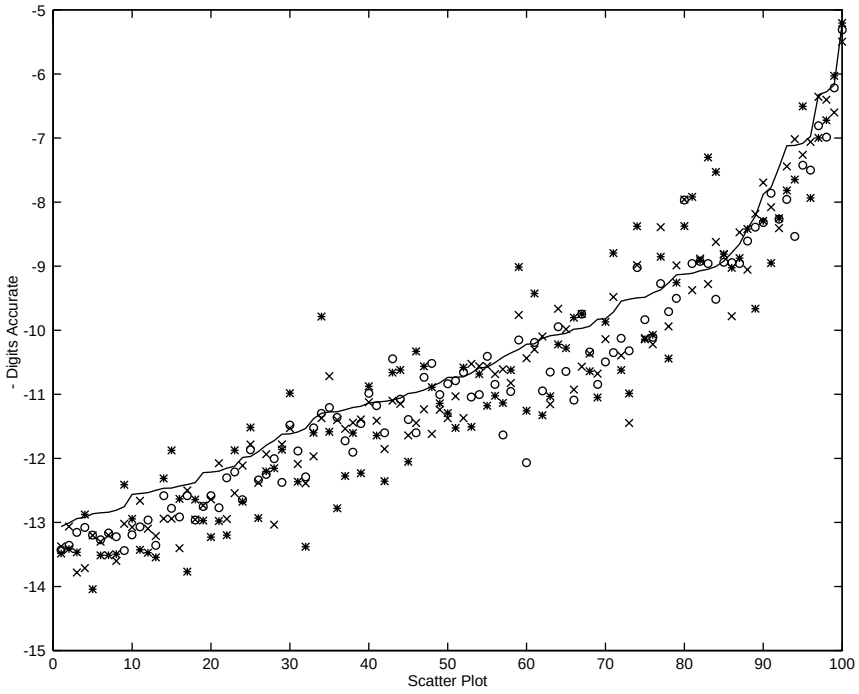


FIG. 2. *Scatter plot for* 100 *problems of size* 20. *$log_{10}(e_c)$ is (the negative of) the number of correct decimal digits in the computed feedback. The predicted error is given by the "continuous" curve, Datta's method is represented by "*", MP by "o", and the RQ method by "x".*

of correct decimal digits. The least accurate result in the sample is reported under "Minimum" accurate digits, and the average number of correct digits in the sample is reported under "Average". In an attempt to remove the "bias" of conditioning from the statistics, a backward error statistic is also computed as $e_b = e_f/\nu_{d\phi}$. The justification for this statistic is that given a backward stable method, the *true condition number* ν, and a small (relative to $1/\nu$) machine epsilon μ, the quantity e_f/ν should be approximately bounded by μ. Thus, we define the quantity "backward digits accurate" as $-\log_{10}(e_b)$. The least accurate sample with respect to this scaled error is reported as "Backward minimum," and the average of the scaled errors is reported as "Backward average."

The next experiment is constructed as the first, but with $n = 20$, and with ill conditioning introduced by uniformly scaling the subdiagonal entries of H so that the product of these entries is between 1×10^{-10} and unity. Again, we include a scatter plot for 100 runs, and a table summarizing the results; these are given in Figure 2 and Table 2, respectively.

TABLE 2
Summary statistics of relative errors for 100 *randomly generated systems of order* 20 *of varying degrees of ill conditioning.*

	Accurate Digits			
Method	*Minimum*	*Average*	*Backward average*	*Backward minimum*
M&P	5.31	10.8	16.0	14.5
RQ	5.50	10.8	16.0	14.5
Datta	5.21	10.8	16.0	13.9

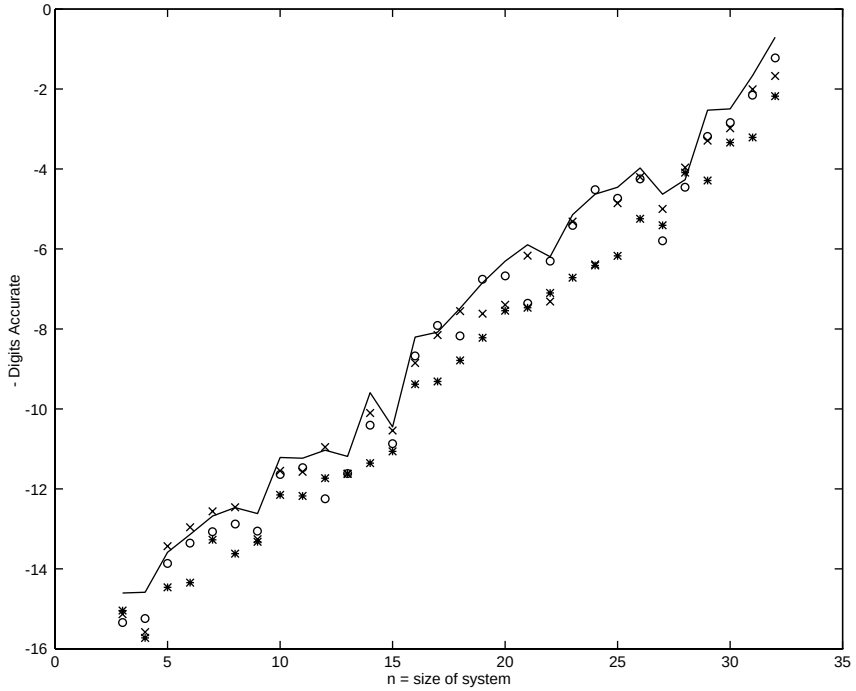


FIG. 3. *Plot for* 29 *problems of size* $n = 3$ *to* $n = 32$. *Now the x-axis represents the size of the system, and the data has not been sorted.* $log_{10}(e_c)$ *is (the negative of) the number of correct decimal digits in the computed feedback. The predicted error is given by the "continuous" curve, Datta's method is represented by "*", MP by "o", and the RQ method by "x".*

The last experiment is constructed as the first, but with the Hessenberg matrix H always set to (see [28])

$$H = \begin{bmatrix} -1 & -1 & -1 & \cdots & -1 \\ 1 & -1 & -1 & \cdots & -1 \\ 0 & 1 & -1 & \cdots & -1 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \cdots & 0 & 1 & 1 \end{bmatrix}.$$

A perturbation on the order of 2^{1-n} makes this system uncontrollable. The system size varied from $n = 3$ to $n = 32$, and one sample was taken for each n. The results are reported in Figure 3 and Table 3. In Figure 3, we display the errors as a function of n, and as such, the data are not sorted.

TABLE 3
Summary statistics of relative errors for Example 3 *of orders* 3 *to* 32.

	Accurate Digits			
Method	*Average*	*Minimum*	*Backward average*	*Backward minimum*
M&P	8.52	1.22	16.1	15.5
RQ	8.48	1.67	16.1	15.4
Datta	9.16	2.18	16.7	15.5

Summary and conclusions. In this paper, we have considered various computational aspects of the single-input eigenvalue assignment problem in control theory. We summarize the results of the paper below.

I. We have built a framework around which the QR-based methods are all special cases. We have found that these apparently different methods differ only on how the RQ decompositions are computed.
II. We have proposed a new method based on the RQ formulation of the recursive algorithm of Datta [17]. An intimate relationship of the latter with the other QR methods has been exposed via an explicit formula of the feedback vector obtained from the recursive algorithm.
III. We have proved that the proposed algorithm is backward stable by a round-off error analysis. A more general theorem obtained in this context shows that an algorithm is backward stable if the associated Hessenberg algorithm is stable. It remains to be seen if the stability of the other QR algorithms can be reproved from the relationship mentioned in I.
IV. We have given a detailed round-off error analysis of the recursive algorithm. Our analysis shows precisely when the results are suspect, and this phenomenon can be determined as the algorithm proceeds, in a relatively inexpensive way.
V. We have reported the results of a numerical comparison of some of the methods.

In the multi-input case, even though the solution is not unique, it might still be possible to obtain a relationship between the solutions obtained by different algorithms. A parameterized expression for the closed-form solution obtained in the thesis of Arnold [2] might play an important role in this context. Also, a QR formulation of the multi-input algorithm in Arnold and Datta [3] is in order.

Acknowledgment. We gratefully acknowledge the constructive remarks made by one of the referees. This careful reading and insightful review has improved the paper.

REFERENCES

[1] J. ACKERMANN, *Der Entwurf Linear Regelungssysteme im Zustandsraum*, Regulungestechnik und Prozessedatenverarbeitung, 7 (1972), pp. 297–300.
[2] M. ARNOLD, *Algorithms and Conditioning for Eigenvalue Assignment,* Ph.D. dissertation, Northern Illinois University, DeKalb, IL, 1993.
[3] M. ARNOLD AND B. N. DATTA, *An algorithm for the multiinput eigenvalue assignment problem*, IEEE Trans. Automat. Control, 35 (1990), pp. 1149–1152.
[4] S. P. BHATTACHARYYA AND E. DESOUZA, *Pole assignment via Sylvester's equation*, Systems Control Lett., 1 (1982), pp. 261–283.
[5] D. H. BAILEY, *A Portable High Performance Multiprecision Package*, RNR Technical report RNR-90-022, NASA Ames Research Center, Moffelt Field, CA, 1992.

[6] D. H. Bailey, *Automatic Translation of Fortran Programs to Multiprecision*, RNR Technical report RNR-91-025, NASA Ames Research Center, Moffelt Field, CA, 1992.
[7] D. L. Boley, *Computing the Controllability/Observability Decomposition of a Linear Time-Invariant Dynamic System, A Numerical Approach*, Ph.D. thesis, Report STAN-CS-81-860, Dept. of Computer Science, Stanford University, Stanford, CA, 1981.
[8] R. Bru, J. Mas, and A. Urbano, *An algorithm for the single input pole assignment problem*, SIAM J. Matrix Appl., 15 (1994), pp. 393–407.
[9] R. Bru, J. Cerdan, and A. Urbano, *An algorithm for the multi-input pole assignment problem*, Linear Algebra Appl., 199 (1994b), pp. 427–444.
[10] R. Bru, J. Cerdan, P. Fernandez de Cordoba, and A. Urbano, *A parallel algorithm for the partial single-input pole assignment problem*, Appl. Math. Lett., 7 (1994), pp. 7–11.
[11] C. T. Chen, *Linear System Theory and Design*, CBS College Publishing, New York, 1984.
[12] E. Chu and B. N. Datta, *Numerically robust pole assignment for second-order systems*, Internat. J. Control, 64 (1996), pp. 1113–1127.
[13] 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.
[14] C. L. Cox and W. F. Moss, *Backward error analysis of a pole assignment algorithm* II*: The complex case*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1159–1171.
[15] M. G. Coutinho, A. Bhaya, and B. N. Datta, *Parallel algorithm for the eigenvalue assignment problem in linear systems*, in Proc. Int. Conference on Control and Information, Hong Kong, 1995, pp. 163–168.
[16] B. N. Datta and K. Datta, *Efficient parallel algorithms for controllability and eigenvalue assignment problems*, in Proc. 25th IEEE Conference on Decision and Control, Athens, Greece, 1986, pp. 1611–1616.
[17] B. N. Datta, *An algorithm to assign eigenvalues in a Hessenberg matrix: Single input case*, IEEE Trans. Automat. Control, AC-32 (1987), pp. 414–417.
[18] B. N. Datta and K. Datta, *On eigenvalue and canonical form assignments*, Linear Algebra Appl., 131 (1990), pp. 161–182.
[19] B. N. Datta, *Parallel algorithms in control theory*, in Proc. IEEE Conference on Decision and Control, Birmingham, England, 1991, pp. 1700–1704.
[20] B. N. Datta, *Numerical Algorithms for the Eigenvalue Assignment Problem via Observer Matrix Equation*, in Proc. IEEE Conference on Decision and Control, HI, 1990.
[21] B. N. Datta, *Numerical Linear Algebra and Applications,* Brooks/Cole, Pacific Grove, CA, 1995.
[22] B. N. Datta, S. Elhay, and Y. Ram, *Orthogonality and partial pole assignment for the symmetric quadratic definite pencil*, Linear Algebra Appl., 257 (1997), pp. 29–48.
[23] B. N. Datta and Y. Saad, *Arnoldi methods for large Sylvester-like matrix equations and an associated algorithm for partial spectrum assignment*, Linear Algebra Appl., 154–156 (1991), pp. 225–244.
[24] T. Kailath, *Linear Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[25] J. Kautsky, N. K. Nichols, and P. Van Dooren, *Robust pole assignment in linear state feedback*, Internat. J. Control, 41 (1985), pp. 1129–1155.
[26] The MathWorks, Inc., *MATLAB User's Guide*, The MathWorks, Inc., South Natick, MA, 1992.
[27] 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.
[28] G. S. Miminis, *Numerical Algorithms for Controllability and Eigenvalue Allocation*, M.Sc. thesis, School of Computer Science, McGill University, Montreal, 1981.
[29] G. S. Miminis and C. C. Paige, *A Direct algorithm for pole assignment of time-invariant multi-input linear systems using state feedback*, Automatica, 24 (1988), pp. 343–356.
[30] G. S. Miminis, *Polepack*, (A collection of Matlab programs for Eigenvalue Assignment, available on netlib from www.netlib.org), 1991.
[31] C. C. Paige, *Properties of numerical algorithms related to computing controllability*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 130–138.
[32] R. V. Patel and P. Misra, *Numerical algorithms for eigenvalue assignment by state feedback*, Proc. IEEE, 72 (1984), pp. 1755–1764.
[33] P. Hr. Petkov, N. D. Christov, and M. M. Konstantinov, *A computational algorithm for pole assignment of linear multiinput systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 1044–1047.
[34] 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.

[35] Y. Saad, *Projection and deflation method for partial pole assignment in linear state feedback*, IEEE Trans. Automat. Control, AC-33 (1988), pp. 290–297.
[36] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, Orlando, FL, 1973.
[37] C. C. Tsui, *An algorithm for computing state feedback in multiinput linear systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 243–246.
[38] F. Szidarovszky and A. T. Bahil, *Linear Systems Theory*, CRC Press, Boca Raton, FL, 1991.
[39] P. M. Van Dooren and M. Verhaegen, *On the use of unitary state-space transformations*, Contemp. Math., 47 (1985), pp. 447–463.
[40] A. Varga, *A Schur method for pole assignment*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 517–519.
[41] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.
[42] W. M. Wonham, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer, New York, 1979.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 468–486, April 1998

012

A CUBICALLY CONVERGENT PARALLELIZABLE METHOD FOR THE HERMITIAN EIGENVALUE PROBLEM*

HONGYUAN ZHA† AND ZHENYUE ZHANG‡

Abstract. We propose a *cubically* convergent algorithm for computing the invariant subspaces of an Hermitian matrix. The building blocks of the algorithm are matrix–matrix multiplication and QR decomposition which are highly parallelizable. We present a detailed convergence analysis and explore the so-called *mixed convergence* phenomenon, the understanding of which will be very helpful in devising convergence improvement. We also discuss a number of implementation details and demonstrate convergence properties of the algorithm using several numerical examples.

Key words. invariant subspace, matrix power, Hermitian eigenproblem, cubic convergence

AMS subject classifications. 65F05, 65F35

PII. S0895479896302035

1. Introduction.

The Hermitian eigenproblem is one of the fundamental problems in matrix computations. A number of methods have been proposed in the past for computing the eigenvalues and eigenvectors of a dense Hermitian matrix, notably, the QR method, Cuppen's divide-and-conquer method, and Jacobi method [4, Chapter 8]. Recently, there has been much interest in developing algorithms for the Hermitian and/or general eigenproblems that can be efficiently implemented on a variety of parallel computers [1, 2, 3, 6, 7, 8]. Those algorithms are iterative in nature and rely on a few operations such as matrix–matrix multiplication, QR decomposition, and matrix inversion as their building blocks at each iteration. Those operations have already been successfully implemented on many parallel architectures, and portable parallel library is now available for their efficient computation [9].

In this paper, we are concerned with the problem of computing the invariant subspaces of an Hermitian matrix A corresponding to eigenvalues inside or outside the interval $(-1, 1)$.[1] We use $P_{|\lambda|<1}$ and $P_{|\lambda|>1}$ to denote the orthogonal projections onto these two invariant subspaces. Other intervals on the real line can also be considered by using a suitable linear transformation. An elegant method for computing those invariant subspaces is the matrix sign function method, a simple scheme of which is the following iteration:[2]

$$A_{j+1} = (A_j + A_j^{-1})/2, \quad j = 0, 1, \ldots, \quad A_0 = A.$$

However, as pointed out in [2], numerical difficulty will occur when A is ill conditioned with respect to inversion. Another problem is that without explicitly forming the

*Received by the editors April 15, 1996; accepted for publication (in revised form) by A. Edelman April 8, 1997.

http://www.siam.org/journals/simax/19-2/30203.html

†Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802 (zha@cse.psu.edu). The work of this author was supported in part by NSF grants CCR-9308399 and CCR-9619452.

‡Center for Mathematical Sciences and Department of Applied Mathematics, Zhejiang University, Hangzhou, 310027, P. R. China, and Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802 (zyzhang@cse.psu.edu). The work of this author was supported in part by China State Major Key Project for Basic Researches and by the Board of Pao Yu-Kong and Pao Zhao-Long Scholarship for Chinese Students Studying Abroad.

[1] Assume for the moment that A has no eigenvalue that is either 1 or -1.

[2] This iteration computes the invariant subspaces corresponding to eigenvalues less than and greater than zero, respectively. The convergence rate is quadratic.

product $A^H A$, the matrix sign function method cannot readily be extended to the case when singular subspaces of a general matrix A are desirable. For example, the above iteration scheme converges to the unitary factor in the polar decomposition of A [5]. The major inspiration for the work reported in this paper came from [8, 2], where a *quadratically* convergent inverse-free method was proposed for computing the invariant subspaces of a general matrix. In fact the work was started when we tried to simplify the algorithms in [8, 2] for the Hermitian case. Another inspiration is the Project PRISM reported in [3, 7] which forced us to think more carefully about a variety of implementation issues. A number of issues will be addressed in this paper. We propose a *cubically* convergent algorithm for computing the orthogonal projections $P_{|\lambda|<1}$ and $P_{|\lambda|>1}$. We will also present a detailed and rigorous convergence analysis of the proposed algorithm. We observed a phenomenon which we call *mixed convergence* that will have practical important consequences in the implementations of the algorithm. We show that certain quantities in our algorithm can be used to overcome mixed convergence, and therefore accelerates the convergence of the algorithm. Here is an outline of the rest of the paper. In section 2, we start with a very intuitive idea based on computing the successive powers of A to find approximations of $P_{|\lambda|<1}$ and $P_{|\lambda|>1}$, and we also provide a formal analysis. Concerns about the possible overflow of A^m leads to an improved method which we call the basic iterative scheme. In section 3, we propose an algorithm based on several refinements of the basic iterative scheme. Section 4 is devoted to a detailed convergence analysis of the algorithm, and an exploration of the *mixed convergence* phenomenon. In section 5, we discuss several implementation details, and we present several numerical examples in section 6. We conclude the paper with some remarks about future research in section 7.

Notation. Let the eigenvalues of the Hermitian matrix $A \in \mathcal{C}^{n\times n}$ be $\{\lambda_1, \ldots, \lambda_n\}$, and

$$|\lambda_i| < 1,\ 1 \le i \le r, \quad |\lambda_i| = 1,\ r < i \le s, \quad |\lambda_i| > 1,\ s < i \le n. \tag{1.1}$$

If $r = s$, then A has no eigenvalue of absolute value one. We define

$$\eta_- = \max\{|\lambda_i| \mid 1 \le i \le r\}, \quad \eta_+ = \max\{|\lambda_i^{-1}| \mid s < i \le n\}, \quad \eta = \max\{\eta_-, \eta_+\}.$$

Then $\eta < 1$. We will also use the $O(\cdot)$ notation, so two sequences $\{x_n\}$ and $\{y_n\}$ $x_n = O(y_n)$ mean $|x_n| \le C|y_n|$ for all integers $n \ge 1$, where C is a constant independent of n. The singular values of a matrix G are denoted by $\sigma_1(G) \ge \cdots \ge \sigma_n(G)$ and we also use $\sigma_{\max}(G) = \sigma_1(G)$ and $\sigma_{\min}(G) = \sigma_n(G)$.

2. A basic iterative scheme and analysis. In this section we will first informally motivate the derivation of a basic iterative scheme by examining the convergence property of the successive powers of A. The attempt to avoid overflow in forming the powers $\{A^m\}$ leads us to the basic iterative scheme. We then present a convergence analysis. In the next section, an algorithm will be proposed for computing the invariant subspaces based on several refinements of the basic iterative scheme.

Let A be an Hermitian matrix and assume that A is block diagonalized as follows:

$$A = Q\,\mathrm{diag}(A_{11}, A_{22})Q^H, \tag{2.1}$$

where Q is unitary, and $A_{11} \in \mathcal{C}^{r\times r}$ with

$$\lambda(A_{11}) = \{\lambda \in \lambda(A) \mid |\lambda| < 1\}, \quad \lambda(A_{22}) = \{\lambda \in \lambda(A) \mid |\lambda| > 1\},$$

i.e., A has no eigenvalue of absolute value one. Our goal is to compute the unitary matrix Q whose first r columns form an orthonormal basis of the invariant subspace of A corresponding to the eigenvalues inside $(-1, 1)$. To this end, it is easily verified that

$$\begin{aligned}(I + A^m)^{-1} &= Q\,\mathrm{diag}((I + A_{11}^m)^{-1}, (I + A_{22}^m)^{-1})Q^H \\ &= Q\,\mathrm{diag}(I_r, 0)Q^H + O(\epsilon_m) \\ &= P_{|\lambda|<1} + O(\epsilon_m),\end{aligned}$$

where $\epsilon_m = \eta^m$; here m is a nonnegative integer. Therefore, an approximation of $P_{|\lambda|<1}$ can be obtained by using a large enough m. In fact let $(I+A^m)^{-1} = Q_m R_m \Pi_m$ be its QR decomposition with column pivoting. We now prove that the $(2, 1)$ block of $Q_m^H A Q_m$ converges to zero, and we will also establish its convergence rate. Before we prove the convergence result, we need the following lemma.

LEMMA 2.1. *Let $V \in \mathcal{C}^{r\times n}$ satisfy $VV^H = I_r$, and $r < n$. Let $V = QR\Pi$ be the QR decomposition of V with column pivoting. Partition R as $R = (R_1, R_2)$ with $R_1 \in \mathcal{C}^{r\times r}$. Then*

$$\sigma_{\min}^2(R_1) \geq 1/C_n^r,$$

where $C_n^r = n!/(r!(n-r)!)$.

Proof. Since $\sigma_{\max}(R_1) \leq 1$, we have

$$\sigma_{\min}(R_1) \geq \sigma_1(R_1)\cdots\sigma_r(R_1) = \det(R_1).$$

Let $R = [r_1, \ldots, r_n]$. It follows that $r = \|R\|_F^2 \leq n \max\{\ \|r_j\|_2^2 \ \mid\ 1 \leq j \leq n\}$. Let $R_1 = (R_{ij})_{i,j=1}^r$. Now R_{11} is bounded below by

$$R_{11} = \max\{\ \|r_j\|_2 \ \mid\ 1 \leq j \leq n\} \geq \sqrt{r/n}.$$

Similarly we have

$$R_{jj} \geq \sqrt{(r-j+1)/(n-j+1)}, \quad j = 1, 2, \ldots, r.$$

Hence

$$\sigma_{\min}^2(R_1) \geq (\det(R_1))^2 = \prod_{j=1}^r R_{jj}^2 \geq \frac{r(r-1)\cdots 2\cdot 1}{n(n-1)\cdots(n-r+1)} = \frac{1}{C_n^r},$$

completing the proof. ∎

THEOREM 2.2. *Let $(I+A^m)^{-1} = Q_m R_m \Pi_m$ be its QR decomposition with column pivoting. Partition*

$$Q_m^H A Q_m = \begin{bmatrix} A_{11}^{(m)} & A_{12}^{(m)} \\ A_{21}^{(m)} & A_{22}^{(m)} \end{bmatrix},$$

with $A_{11}^{(m)} \in \mathcal{C}^{r\times r}$. Then we have

$$A_{12}^{(m)} = (A_{21}^{(m)})^H = O(\epsilon_m)$$

with $\epsilon_m = \eta^m$.

Proof. First let $P_m = \Pi_m Q, U_m = Q_m^H Q$. Then we have

$$R_m P_m = U_m \operatorname{diag}(I_r, 0) + O(\epsilon_m), \tag{2.2}$$

which gives

$$R_m = U_m \operatorname{diag}(I_r, 0) P_m^H + O(\epsilon_m). \tag{2.3}$$

Now partition R_m, P_m, and U_m conformally as follows:

$$R_m = \begin{bmatrix} R_{11}^{(m)} & R_{12}^{(m)} \\ R_{21}^{(m)} & R_{22}^{(m)} \end{bmatrix}, \quad P_m = \begin{bmatrix} P_{11}^{(m)} & P_{12}^{(m)} \\ P_{21}^{(m)} & P_{22}^{(m)} \end{bmatrix}, \quad U_m = \begin{bmatrix} U_{11}^{(m)} & U_{12}^{(m)} \\ U_{21}^{(m)} & U_{22}^{(m)} \end{bmatrix}$$

with $R_{11}^{(m)} \in \mathcal{C}^{r\times r}$. It follows from (2.2) and (2.3) that

$$U_{21}^{(m)} = R_{22}^{(m)} P_{21}^{(m)} + O(\epsilon_m), \quad R_{22}^{(m)} = U_{21}^{(m)} (P_{21}^{(m)})^H + O(\epsilon_m), \tag{2.4}$$

which implies

$$U_{21}^{(m)} (I - (P_{21}^{(m)})^H P_{21}^{(m)}) = O(\epsilon_m).$$

Therefore, $U_{21}^{(m)} (P_{11}^{(m)})^H P_{11}^{(m)} = O(\epsilon_m)$.

Now we show that $R_{22}^{(m)} \to 0$ as $m \to \infty$. Since $\{R_{22}^{(m)}\}$ is bounded, this is equivalent to proving that any convergent subsequence of $\{R_{22}^{(m)}\}$ has zero as its limit. To this end, choose a convergent subsequence $\{R_{22}^{(m_k)}\}$ and without loss of generality assume that $\{R_{m_k}\}$, $\{P_{m_k}\}$, and $\{U_{m_k}\}$ are convergent to R, P, and U, respectively. It is easy to see that

$$\operatorname{diag}(I_r, 0) P = U^H R$$

is the QR decomposition with column pivoting of $\operatorname{diag}(I_r, 0)P$. Hence we have $R_{22} = 0$ because $\operatorname{rank}(R) = r$, i.e., $R_{22}^{(m_k)} \to 0$ as $k \to \infty$.

Next we show that $\sigma_{\max}(P_{21}^{(m)}) \le 1 - 1/C_n^r, m = 0, 1, \ldots$. In fact take any convergent subsequence $\{P_{21}^{(m_k)}\}$ of $\{P_{21}^{(m)}\}$ and assume without loss of generality

$$P_{m_k} \to \begin{bmatrix} P_{11} & P_{21} \\ P_{21} & P_{22} \end{bmatrix}, \quad U_{m_k} \to \begin{bmatrix} U_{11} & U_{21} \\ U_{21} & U_{22} \end{bmatrix}, \quad R_{m_k} \to \begin{bmatrix} R_{11} & R_{21} \\ 0 & 0 \end{bmatrix}.$$

Since $R_{22}^{(m_k)} \to 0$, we have from (2.4) that $U_{21} = 0$ and then $U_{12} = 0$. Hence U_{11} is orthogonal. Therefore,

$$(R_{11}, R_{12}) = U_{11} (P_{11}^H, P_{21}^H)$$

is the QR decomposition with column pivoting of (P_{11}^H, P_{21}^H). By Lemma 2.1, we have

$$\sigma_{\min}(P_{11}) = \sigma_{\min}(R_{11}) \ge 1/C_n^r,$$

which implies that $\sigma_{\min}(P_{11}^{(m)}) \ge 1/C_n^r$ because the set $\{P_m\}$ only contains a finite number of elements. ($P_m = \Pi_m Q$, where Π_m is a permutation matrix, and the number of permutation matrices of order n is finite.)

Therefore, we have $(U_{21}^{(m)}, R_{22}^{(m)}) = O(\epsilon_m)$. On the other hand

$$U_m^H R_m = \operatorname{diag}(I_r, 0) P_m^H + O(\epsilon_m),$$

which gives $(U_{12}^{(m)})^H (R_{11}^{(m)}, R_{12}^{(m)}) = O(\epsilon_m)$. Recall that $(R_{11}^{(m)}, R_{12}^{(m)})(R_{11}^{(m)}, R_{12}^{(m)})^H \to I_r$, we have $U_{12}^{(m)} = O(\epsilon_m)$. It follows from

$$A_{21}^{(m)} = U_{21}^{(m)} A_{11} (U_{11}^{(m)})^H + U_{22}^{(m)} A_{22} (U_{12}^{(m)})^H$$

that $A_{21}^{(m)} = O(\epsilon_m)$. □

Remark. Using the estimates $U_{12}^{(m)} = O(\epsilon_m)$ and $U_{21}^{(m)} = O(\epsilon_m)$ in the above proof, we can show that there exist unitary matrices $P_1^{(m)} \in \mathcal{C}^{r\times r}$ and $P_2^{(m)} \in \mathcal{C}^{(n-r)\times(n-r)}$ such that

$$A_{11}^{(m)} = P_1^{(m)} A_{11} (P_1^{(m)})^H + O(\epsilon_m^2), \quad A_{22}^{(m)} = P_2^{(m)} A_{22} (P_2^{(m)})^H + O(\epsilon_m^2).$$

The proof is similar to that of the second part of Theorem 4.5. Theorem 2.2 and the above remark show that the first r columns of Q_m span an approximate orthonormal basis for the invariant subspace of A corresponding to its eigenvalues *inside* $(-1, 1)$, and the last $n - r$ columns of Q_m span an approximate orthonormal basis for the invariant subspace of A corresponding to its eigenvalues *outside* $(-1, 1)$.

Unfortunately, the above method cannot be used as it is because forming A^m for large m will result in overflow. As a remedy, we can rewrite

$$A^m = Q \operatorname{diag}(I, A_{22}^{-m})^{-1} Q^H Q \operatorname{diag}(A_{11}^m, I) Q^H \equiv W_m^{-1} Z_m,$$

where

$$W_m = Q \operatorname{diag}(I, A_{22}^{-m}) Q^H, \quad Z_m = Q \operatorname{diag}(A_{11}^m, I) Q^H.$$

The computation of W_m and Z_m now causes no overflow, and we also have

$$(I + A^m)^{-1} = (I + W_m^{-1} Z_m)^{-1} = (W_m + Z_m)^{-1} W_m.$$

To speed up convergence, we will not generate the successive powers $\{(I+A^m)^{-1}\}$ one by one. Instead we will just generate $(I + A^m)^{-1}$ for $m = 3^k, k = 0, 1, \ldots$. The following iterative scheme does this.

A basic iterative scheme.

1. *Initialization.* $W_0 = I$, $Z_0 = A$.
2. For $k = 0, 1, 2, \ldots$, until convergence
   2.1 Compute QR decomposition

$$\begin{bmatrix} W_k \\ -Z_k \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R_k \\ 0 \end{bmatrix}.$$

   2.2 Modify W_k and Z_k

$$W_{k+1} = Q_{22}^H W_k Q_{22},$$
$$Z_{k+1} = Q_{12}^H Z_k Q_{12}.$$

Notice that step 2.1 amounts to the computation of an orthonormal basis for the *left* null space of $(W_k^H, -Z_k^H)^H$, i.e.,

$$[Q_{12}^H, Q_{22}^H] \begin{bmatrix} W_k \\ -Z_k \end{bmatrix} = 0. \tag{2.5}$$

The connections between the basic iterative scheme and the powers of $\{A^m\}$ are given in the following results.

PROPOSITION 2.3. *The matrices W_k and $W_k + Z_k$, $k = 0, 1, \ldots$, are all nonsingular, and there exists a nonsingular matrix G_k such that*

$$W_k^{-1} Z_k = G_k^{-1} A^{3^k} G_k, \quad (I + A^{3^k})^{-1} = G_k (W_k + Z_k)^{-1} G_k^H.$$

Proof. First we prove the nonsingularity of W_k by induction. It is trivial for $k = 0$ since $W_0 = I$. Assume that W_k is nonsingular; then from

$$\begin{bmatrix} W_k \\ -Z_k \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R_k \\ 0 \end{bmatrix}$$

we obtain $W_k = Q_{11} R_k$ which implies that Q_{11} is also nonsingular. We prove W_{k+1} is nonsingular by proving Q_{22} is so. Assume to the contrary that Q_{22} is singular, and $Q_{22} x = 0$ with $x \neq 0$. Since Q is unitary, we have $Q_{11}^H Q_{12} x = 0$. Hence $Q_{12} x = 0$, and

$$\begin{bmatrix} Q_{12} \\ Q_{22} \end{bmatrix} x = 0,$$

a contradiction. At each step 2.1 in the algorithm, let $X_k = Q_{22}^H$ and $Y_k = Q_{12}^H$. It follows from (2.5) that $Y_k W_k = X_k Z_k$, and

$$W_{k+1} = X_k W_k X_k^H, \quad Z_{k+1} = Y_k Z_k Y_k^H.$$

Using induction, it is easy to see that W_k and Z_k, $k = 0, 1, \ldots$, are all Hermitian, and $W_k = G_k^H G_k$ with $G_k = (X_{k-1} \ldots X_0)^H$. Now

$$\begin{aligned} W_{k+1}^{-1} Z_{k+1} &= X_k^{-H} W_k^{-1} X_k^{-1} (Y_k) Z_k (Y_k^H) \\ &= X_k^{-H} W_k^{-1} X_k^{-1} (X_k Z_k W_k^{-1}) Z_k (W_k^{-1} Z_k X_k^H) \\ &= X_k^{-H} (W_k^{-1} Z_k)^3 X_k^H, \end{aligned}$$

which leads to

$$W_k^{-1} Z_k = G_k^{-1} (W_0^{-1} Z_0)^{3^k} G_k = G_k^{-1} A^{3^k} G_k.$$

Moreover, with $W_k = G_k^H G_k$, we obtain

$$I + A^{3^k} = G_k^{-H} (W_k + Z_k) G_k.$$

Therefore, $W_k + Z_k$ is nonsingular, and

$$(I + A^{3^k})^{-1} = G_k (W_k + Z_k)^{-1} G_k^H,$$

completing the proof. ∎

With the relation $(I + A^{3^k})^{-1} = G_k(W_k + Z_k)^{-1}G_k^H$, it is just a matter of computing the QR decomposition of $G_k(W_k + Z_k)^{-1}G_k^H$ with column pivoting to find an approximation of $P_{|\lambda|>1}$. Let this decomposition be $G_k(W_k+Z_k)^{-1}G_k^H = Q_k R_k \Pi_k$. A simple corollary of Theorem 2.2 is this: if Q_k is the unitary matrix in the QR decomposition of $G_k(W_k + Z_k)^{-1}G_k^H$ with column pivoting, then the $(2,1)$ block of $Q_k^H A Q_k$ is of order η^{3^k}. This establishes its cubic convergence. However, the computation of $(W_k + Z_k)^{-1}$ can introduce large errors if $W_k + Z_k$ is close to singular. Besides, we are also interested in designing methods in which no inversion of any matrices is involved which will give us truly *inverse-free* algorithms. This task is taken up in the next section. The convergence analysis in Theorem 2.2 can be best described as using a mixed convergence rate because η is actually the maximum of η_+ and η_-. A more desirable feature will be to identify quantities in the iterative scheme that will have convergence rates determined by either η_+ or η_-. It is easy to see that there are cases where η_+ can be much smaller than η_- or vice versa.

3. The algorithm. In this section we will present an algorithm based on some refinements of the basic iterative scheme proposed in the last section. We will also give a much refined convergence analysis where convergence to $P_{|\lambda|>1}$ and $P_{|\lambda|<1}$ are clearly separated with distinct convergence rates.

A key in the refinement of the basic iterative scheme is the observation that the matrices W_k and Z_k, $k = 0, 1, \ldots$, in the basic iterative scheme actually commute with each other, i.e., $W_k Z_k = Z_k W_k$. This result is proved in Proposition 3.1 below, and it allows us to replace the computation of an orthonormal basis for the *left* null space of $(W_k^H, -Z_k^H)^H$ (cf. equation (2.5)) by the computation of an orthonormal basis for the range space of $(Z_k^H, W_k^H)^H$. Moreover, instead of directly generating the matrices W_k and Z_k, we actually generate G_k and H_k such that $W_k = G_k^H G_k$ and $Z_k = H_k^H A H_k$. Other alternatives are possible and will be discussed in section 5.

PROPOSITION 3.1. *Let* $[(Q_{12}^{(k)})^H, (Q_{22}^{(k)})^H]$ *be any orthonormal matrix in step* 2.1 *of the basic iterative scheme such that*

$$[(Q_{12}^{(k)})^H, (Q_{22}^{(k)})^H] \begin{bmatrix} W_k \\ -Z_k \end{bmatrix} = 0.$$

Let $G_{k+1} = G_k(Q_{22}^{(k)})^H, H_{k+1} = H_k(Q_{21}^{(k)})^H$ *with* $G_0 = H_0 = I$. *Then there exists a unitary matrix* M_k *such that for* $k \geq 1$

$$G_k = C_k M_k, \quad H_k = A^{(3^k-1)/2} C_k M_k,$$

where $C_k = [(I + A^2)(I + A^{2\cdot 3}) \cdots (I + A^{2\cdot 3^{k-1}})]^{-1/2}$. *Furthermore,* W_k *and* Z_k *commute with each other.*

Proof. We use induction. At the first step of the basic iterative scheme, we have

$$\begin{bmatrix} I \\ -A \end{bmatrix} = \begin{bmatrix} Q_{11}^{(0)} & Q_{12}^{(0)} \\ Q_{21}^{(0)} & Q_{22}^{(0)} \end{bmatrix} \begin{bmatrix} R_0 \\ 0 \end{bmatrix},$$

the QR decomposition of $(I, -A)^H$. We also know R_0 is nonsingular. It is easy to see that

$$\begin{bmatrix} A \\ I \end{bmatrix} (I + A^2)^{-1/2} \text{ is orthogonal to } \begin{bmatrix} Q_{11}^{(0)} \\ Q_{21}^{(0)} \end{bmatrix}.$$

Therefore, there exists a unitary matrix M_1 such that

$$\begin{bmatrix} A \\ I \end{bmatrix} (I + A^2)^{-1/2} = \begin{bmatrix} Q_{12}^{(0)} \\ Q_{22}^{(0)} \end{bmatrix} M_1^H.$$

Hence

$$G_1 = Q_{22}^{(0)} = (I + A^2)^{-1/2} M_1, \quad H_1 = Q_{12}^{(0)} = A(I + A^2)^{-1/2} M_1,$$

and the proposition holds for $k = 1$. Now assume that the result holds for $k = m \geq 1$. The QR decomposition of $(W_m^H, -Z_m^H)^H$ gives

$$\begin{bmatrix} W_m \\ -Z_m \end{bmatrix} \equiv \begin{bmatrix} G_m^H G_m \\ H_m^H A H_m \end{bmatrix} = \begin{bmatrix} M_m^H C_m^2 M_m \\ M_m^H A^{3^m - 1} C_m^2 M_m \end{bmatrix} = \begin{bmatrix} Q_{11}^{(m)} & Q_{12}^{(m)} \\ Q_{21}^{(m)} & Q_{22}^{(m)} \end{bmatrix} \begin{bmatrix} R_m \\ 0 \end{bmatrix}.$$

It can be verified that

$$\begin{bmatrix} M_m^H A^{3^m} \\ M_m^H \end{bmatrix} (I + A^{2 \cdot 3^m})^{-1/2} \text{ is orthogonal to } \begin{bmatrix} Q_{11}^{(m)} \\ Q_{21}^{(m)} \end{bmatrix}.$$

Therefore, there exists a unitary matrix M_{m+1} such that

$$\begin{bmatrix} M_m^H A^{3^m} \\ M_m^H \end{bmatrix} (I + A^{2 \cdot 3^m})^{-1/2} = \begin{bmatrix} Q_{12}^{(m)} \\ Q_{22}^{(m)} \end{bmatrix} M_{m+1}^H.$$

Hence

$$Q_{12}^{(m)} = M_m^H (I + A^{2 \cdot 3^m})^{-1/2} M_{m+1}, \quad Q_{22}^{(m)} = M_m^H A^{3^m} (I + A^{2 \cdot 3^m})^{-1/2} M_{m+1}.$$

Therefore, we have

$$G_{m+1} = G_m Q_{22}^{(m)} = C_{m+1} M_{m+1}, \quad H_{m+1} = H_m Q_{21}^{(m)} = A^{(3^{(m+1)} - 1)/2} C_{m+1} M_{m+1}.$$

Since $W_k = G_k^H G_k = M_m^H C_m^2 M_m$ and $Z_k = H_k^H A H_k = M_m^H A^{3^m - 1} C_m^2 M_m$, it's easy to see that W_k and Z_k commute with each other. □

With the results established in the above proposition, we can present the following algorithm which modifies the basic iterative scheme using the fact that W_k and Z_k commute with each other.

Algorithm Cubic.

1. *Initialization.* $G_0 = I$, $H_0 = I$.
2. For $k = 0, 1, 2, \ldots,$ until convergence
   2.1 Compute QR decomposition
   $$\begin{bmatrix} H_k^H A H_k \\ G_k^H G_k \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R_k \\ 0 \end{bmatrix}.$$
   2.2 Modify G_k and H_k
   $$G_{k+1} = G_k Q_{21},$$
   $$H_{k+1} = H_k Q_{11}.$$
   2.3 If $\|G_{k+1} H_{k+1}^H\|_F \leq \texttt{tol}$, then stop, $p = k + 1$.
   Otherwise $k = k + 1$, go to step 2.1.
3. Compute the QR decomposition with column pivoting
   $$G_p = QR\Pi,$$
   with Π a permutation matrix. Let $r = \text{rank}(R)$.
4. Compute $Q^H A Q$ as
   $$Q^H A Q = \begin{bmatrix} \hat{A}_{11} & E_{12} \\ E_{21} & \hat{A}_{22} \end{bmatrix},$$
   with $\hat{A}_{11} \in \mathcal{C}^{r \times r}$.

Here `tol` is a user-supplied tolerance which will influence the accuracy of the approximate invariant subspaces. More details on the implementation are discussed in section 5.

4. Convergence analysis. In this section we present a detailed convergence analysis of the algorithm proposed in section 3. The analysis is based on several key relations established in Proposition 3.1. We want to emphasize that the cubic convergence of Algorithm Cubic comes from matrix powers, while that of Rayleigh quotient iteration and QR algorithms comes from the choice of shifts. Even though both classes of algorithms converge cubically, the mechanism by which the convergence rate is achieved is entirely different. Now let the eigenvalue decomposition of A be $A = Q\Lambda Q^H$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ satisfying (1.1). Then it is easy to see that C_k defined in Proposition 3.1 has the form $C_k = Q\Lambda_k Q^H$, where $\Lambda_k = \text{diag}(\lambda_1^{(k)}, \ldots, \lambda_n^{(k)})$ and

$$\lambda_i^{(k)} = [(1 + \lambda_i^2)(1 + \lambda_i^{2\cdot 3}) \ldots (1 + \lambda_i^{2\cdot 3^{(k-1)}})]^{-1/2}, \quad i = 1, \ldots, n. \tag{4.1}$$

The following lemma establishes that all the sequences $\{\lambda_i^{(k)}\}, i = 1, \ldots, n,$ converge.

LEMMA 4.1. *Let $d \geq 0$, and define a sequence $\{a_k\}$ by the following recursion:*

$$\begin{aligned} a_1 &= (1+d)^{-1/2} \\ a_{k+1} &= a_k(1+d^{3^k})^{-1/2},\ k \geq 1. \end{aligned}$$

Then the sequence $\{a_k\}$ converges to $a(d)$, where $a(d)$ is a well-defined function of d satisfying $a(d) = 0$ for $d \geq 1$, and $0 < a(d) < 1$ for $0 \leq d < 1$. Moreover, we have

$$|a_k - a(d)| \leq \begin{cases} \frac{a_1}{2(1-d^3)} d^{3^k}, & d < 1, \\ 2^{-k/2}, & d = 1, \\ \sqrt{\frac{d}{d+1}} d^{-(3^k-1)/4}, & d > 1. \end{cases}$$

Proof. It is easy to see that $a_k \to 0$ as $k \to \infty$ when $d \geq 1$. We only need to consider the case when $0 \leq d < 1$. By L'Hospital's rule, we can establish that

$$\lim_{t \to +\infty} t^2 \ln(1 + d^{3^t}) = 0,$$

which implies that there exists a positive constant C such that

$$0 \leq \ln(1 + d^{3^k}) \leq C/k^2$$

holds for all $k \geq 1$. Hence the series $\sum_{k=0}^{\infty} \ln(1 + d^{3^k})$ converges. Let $E \geq 0$ be the sum of the series. Then

$$\lim_{k \to \infty} a_k = \exp(-E/2) \leq 1.$$

As for the upper bounds, we have if $d < 1$, then all $k \geq 1$,

$$|a_{k+1} - a_k| = a_k \frac{d^{3^k}}{1 + d^{3^k} + \sqrt{1 + d^{3^k}}} \leq a_1 d^{3^k}/2,$$

which gives, for $m \geq k$,

$$|a_{m+1} - a_k| \leq a_1(d^{3^k} + d^{3^{(k+1)}} + \cdots + d^{3^m})/2 \leq \frac{a_1}{2(1-d^3)} d^{3^k}.$$

Let $m \to \infty$; we obtain the inequality. If $d = 1$, it is easy to see that $a_k = 2^{-k/2}$. If $d > 1$, we have

$$|a_k - a(d)| = a_k = \frac{d^{-(3^k-1)/4}}{\sqrt{\prod_{j=0}^{k-1}(1 + d^{-3^j})}} \leq \sqrt{\frac{d}{d+1}} d^{-(3^k-1)/4},$$

completing the proof. ∎

Now we are in the position to establish the structure of the limits of the convergent subsequences of $\{G_k\}$ and $\{H_k\}$.

THEOREM 4.2. *Let $A = Q\Lambda Q^H$ be the eigenvalue decomposition of A defined in (1.1). Then for any convergent subsequence $\{G_{k_j}\}$ of $\{G_k\}$, its limit G has the form $G = Q\,\mathrm{diag}(D_G, 0)M$, where M is unitary and*

$$D_G = \mathrm{diag}(a(\lambda_1^2), \ldots, a(\lambda_r^2)).$$

For the corresponding subsequence $\{H_{k_j}\}$ of $\{H_k\}$, the subsequence of $\{H_{k_j}\}$ with odd indices converges to $H = Q\,\mathrm{diag}(0, D_H)M$, where

$$D_H = \mathrm{diag}(\mathrm{sign}(\lambda_{s+1})a(\lambda_{s+1}^{-2}), \ldots, \mathrm{sign}(\lambda_n)a(\lambda_n^{-2})),$$

and the subsequence of $\{H_{k_j}\}$ with even indices converges to $H = Q\,\mathrm{diag}(0, \hat{D}_H)M$, where

$$\hat{D}_H = \mathrm{diag}(a(\lambda_{s+1}^{-2}), \ldots, a(\lambda_n^{-2})).$$

Proof. Equation (4.1) and Lemma 4.1 imply that

$$C_k = [(I + A^2)(I + A^{2\cdot 3}) \cdots (I + A^{2\cdot 3^{k-1}})]^{-1/2}$$

converges to $Q(D_G, 0)Q^H$. Write $A^{(3^k-1)/2}C_k = QD_kQ^H$ with $D_k = \mathrm{diag}(d_1^{(k)}, \ldots, d_n^{(k)})$, and

$$d_i^{(k)} = \lambda_i^{(3^k-1)/2}[(1 + \lambda_i^2)(1 + \lambda_i^{2\cdot 3}) \cdots (1 + \lambda_i^{2\cdot 3^{k-1}})]^{-1/2}.$$

It can be verified that for $1 \le i \le r$, $d_i^{(k)} \to 0 \cdot a(\lambda_i^2) = 0$; for $r < i \le s$, $|d_i^{(k)}| \to a(\lambda_i^2) = 0$; and for $s < i \le n$, we have

$$d_i^{(k)} = \mathrm{sign}(\lambda_i)^k[(I + \lambda_i^{-2})(I + \lambda_i^{(-2)\cdot 3}) \cdots (I + \lambda_i^{(-2)\cdot 3^{k-1}})]^{-1/2}.$$

Hence $d_i^{(2k)} \to a(\lambda_i^{-2})$ and $d_i^{(2k+1)} \to \mathrm{sign}(\lambda_i)a(\lambda_i^{-2})$. Therefore, for any convergent subsequence $\{M_{k_j}\}$ of $\{M_k\}$, the subsequences $\{G_{k_j}\}$ and $\{H_{k_j}\}$ converge to $Q\,\mathrm{diag}(D_G, 0)Q^H\hat{M}$ and $Q\,\mathrm{diag}(0, D_H)Q^H\hat{M}$ or $Q\,\mathrm{diag}(0, \hat{D}_H)Q^H\hat{M}$, provided that $M_{k_j} \to \hat{M}$. Setting $M = Q^H\hat{M}$ finishes the proof. ∎

In Algorithm Cubic, $\|G_kH_k^H\|_F$ is used as a stopping criterion, and $\|G_kH_k^H\|_F \le \sqrt{n}\|G_kH_k^H\|_2$. The following lemma establishes the convergence rate of the sequence $\{\|G_kH_k^H\|_2\}$.

LEMMA 4.3. *For $k \ge 1$, we have $\|G_kH_k^H\|_2 \le \eta^{(3^k-1)/2}$ when $s = r$ and $\|G_kH_k^H\|_2 \le 2^{-k}$ when $s > r$.*

Proof. It is easy to verify that $G_kH_k^H = A^{(3^k-1)/2}C_k^2$. Hence

$$\|G_kH_k^H\|_2 = \max_i \alpha_i^{(k)},$$

where $\alpha_i^{(k)} = |\lambda_i|^{(3^k-1)/2}[(1 + \lambda_i^2)(1 + \lambda_i^{2\cdot 3}) \cdots (1 + \lambda_i^{2\cdot 3^{k-1}})]^{-1}$. Now for $1 \le i \le r$, $\alpha_i^{(k)} \le |\lambda_i|^{(3^k-1)/2} \le \eta_-^{(3^k-1)/2}$; for $r < i \le s$, $\alpha_i^{(k)} = 2^{-k}$; and for $s < i \le n$,

$$\alpha_i^{(k)} = |\lambda_i^{-1}|^{(3^k-1)/2}[(1 + \lambda_i^{-2})(1 + \lambda_i^{(-2)\cdot 3}) \cdots (1 + \lambda_i^{(-2)\cdot 3^{k-1}})]^{-1} \le \eta_+^{(3^k-1)/2},$$

completing the proof. ∎

Before we prove our final result, we need a technical lemma that bounds the diagonal elements of the upper triangular matrices obtained by applying the QR decomposition with column pivoting to a bounded sequence of matrices.

LEMMA 4.4. *If for any convergent subsequence $\{F_{k_j}\}$ of a given bounded matrix sequence $\{F_k\}$ its limit F has rank t, then let $F_k = Q_k R_k \Pi_k$ be its QR decomposition with partial pivoting; we have*

$$\liminf_{k\to\infty} R_{tt}^{(k)} > 0, \quad \lim_{k\to\infty} R_{jj}^{(k)} = 0, \quad j > t,$$

where $R_k = (R_{ij}^{(k)})$.

Proof. Assume to the contrary that $\liminf_{k\to\infty} R_{tt}^{(k)} = 0$. For any convergent subsequence $\{R_{tt}^{k_j}\}$ of $\{F_k\}$, since $\{F_k\}$ is bounded, we can assume without loss of generality that $\{R_{k_j}\}$, $\{Q_{k_j}\}$, and $\{\Pi_{k_j}\}$ are also convergent and have the limits R, Q, and Π, respectively. Then $F = QR\Pi$ is the QR decomposition with column pivoting of F, and $R_{tt} = 0$. Column pivoting guarantees that the last $n - t + 1$ rows of R are also zero which contradicts the assumption that $\operatorname{rank}(R) = t$.

On the other hand, if $R_{t+1,t+1}^{(k)}$ does not converge to zero, then there exits a subsequence $R_{t+1,t+1}^{(k_j)}$ that has a positive limit. Let F be the limit of a convergent subsequence of $F^{(k_j)}$. Then in its QR decomposition with column pivoting $F = QR\Pi$, R has at least $t+1$ nonzero diagonal elements, a contradiction to $\operatorname{rank}(R) = t$. Since $R_{t+1,t+1}^{(k)} \geq R_{jj}^{(k)}, j > t+1$, we also have $R_{jj}^{(k)} \to 0$. ∎

So far, all the results we have obtained about convergence have a convergence rate that is determined by η. Remember that $\eta = \max\{\eta_-, \eta_+\}$. Therefore, the convergence seems to always follow the slowest of η_- and η_+, even though one of them can be much smaller than the other. This will correspond to the situation when either the eigenvalues inside $(-1, 1)$ or the eigenvalues outside $(-1, 1)$ are far from 1 and -1. Using the quantity η alone will not allow us to take advantage of a much smaller η_- or η_+. We call the phenomenon *mixed convergence*. In the following theorem, we will prove that the two sequences $\{G_k\}$ and $\{H_k\}$ have quite different convergence properties: $\{G_k\}$ and $\{H_k\}$ will give a sequence of orthogonal matrices that will block diagonalize A with convergence rate η_- and η_+, respectively.

THEOREM 4.5. *Let $G_k = Q_k R_k \Pi_k$ and $H_k = \hat{Q}_k \hat{R}_k \hat{\Pi}_k$ be QR decomposition with column pivoting for G_k and H_k, respectively. Partition $Q_k^H A Q_k$ and $\hat{Q}_k^H A \hat{Q}_k$ as follows:*

$$Q_k^H A Q_k = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad \hat{Q}_k^H A \hat{Q}_k = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ \hat{A}_{21} & \hat{A}_{22} \end{bmatrix},$$

where $A_{11} \in \mathcal{C}^{r\times r}$ and $\hat{A}_{22} \in \mathcal{C}^{s\times s}$. If $r = s$, then

$$A_{12} = A_{21}^H = O(\eta_+^{3^k/2}), \quad \hat{A}_{12} = \hat{A}_{21}^H = O(\eta_-^{3^k/2}).$$

Moreover, there exist unitary matrices

$$P_1 \in \mathcal{C}^{r\times r}, \quad P_2 \in \mathcal{C}^{(n-r)\times(n-r)}, \quad \hat{P}_1 \in \mathcal{C}^{(n-r)\times(n-r)}, \quad \hat{P}_2 \in \mathcal{C}^{r\times r}$$

such that

$$\begin{aligned} A_{11} - P_1 \Lambda_- P_1^H &= O(\eta_+^{3^k}), \quad A_{22} - P_2 \Lambda_+ P_2^H = O(\eta_+^{3^k}), \\ \hat{A}_{11} - \hat{P}_1 \Lambda_- \hat{P}_1^H &= O(\eta_-^{3^k}), \quad \hat{A}_{22} - \hat{P}_2 \Lambda_+ \hat{P}_2^H = O(\eta_-^{3^k}), \end{aligned} \tag{4.2}$$

where $\Lambda_- = \operatorname{diag}(\lambda_1, \ldots, \lambda_r)$ and $\Lambda_+ = \operatorname{diag}(\lambda_{r+1}, \ldots, \lambda_n)$.

Proof. Denote $U_k = Q_k^H Q$ and $V_k = \hat{Q}_k^H Q$, and partition U_k and V_k as follows:

$$U_k = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}, \quad V_k = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix},$$

where $U_{11} \in \mathcal{C}^{r\times r}$ and $V_{11} \in \mathcal{C}^{(n-r)\times s}$. Since $H_k = A^m G_k$, here $m = (3^k - 1)/2$, we have

$$V_k^H \hat{R}_k \hat{\Pi}_k = \mathrm{diag}(\Lambda_-^m, \Lambda_+^m) U_k^H R_k \Pi_k.$$

Then

$$\mathrm{diag}(I_r, \Lambda_+^{-m}) V_k^H \hat{R}_k \hat{\Pi}_k = \mathrm{diag}(\Lambda_-^m, I_{n-r}) U_k^H R_k \Pi_k.$$

It follows that

$$(V_{11}^H, V_{21}^H)\hat{R}_k = O(\eta_-^m), \quad (U_{12}^H, U_{22}^H) R_k = O(\eta_+^m).$$

Theorem 4.2 and Lemma 4.4 imply that inverses of the leading principal $r \times r$ submatrix of R_k and $(n-r)\times(n-r)$ submatrix of $\hat{R}_k$ have norms with positive lower bounds. Hence $V_{11} = O(\eta_-^m), U_{12} = O(\eta_+^m)$, and then $V_{22} = O(\eta_-^m)$, $U_{21} = O(\eta_+^m)$ which yields $A_{12} = A_{21}^H = O(\eta_+^m)$, and $\hat{A}_{12} = \hat{A}_{21}^H = O(\eta_-^m)$.

Moreover, since $U_{11}^H U_{11} = I_r - U_{21}^H U_{21}$, it follows that there exists a unitary matrix $P_1 \in \mathcal{C}^{r\times r}$ such that $U_{11} = P_1 (I_r - U_{21}^H U_{21})^{1/2}$. Now we have

$$\|U_{11} - P_1\|_2 = \|(I_r - U_{21}^H U_{21})^{1/2} - I_r\|_2 \le \|U_{21}\|_2/2.$$

Using the estimate $U_{21} = O(\eta_+^{3^k/2})$, we have $U_{11} = P_1 + O(\eta_+^{3^k})$. Therefore, we have

$$\begin{aligned} A_{11} - P_1 \Delta_- P_1^H &= U_{11}\Delta_- U_{11}^H + U_{12}\Delta_+ U_{12}^H - P_1 \Delta_- P_1^H \\ &= (U_{11} - P_1)\Delta_- U_{11}^H + P_1 \Delta_- (U_{11} - P_1)^H + U_{12}\Delta_+ U_{12}^H \\ &= O(\eta_+^{3^k}). \end{aligned}$$

Similarly, we can prove the other estimates in (4.2). ∎

Remark. When $r < s$, i.e., A has eigenvalues $|\lambda_i| = 1$, Theorem 4.2 implies that the matrix $Q_k^H A Q_k$ can be written as

$$Q_k^H A Q_k = \begin{bmatrix} A_{11} & E_{12} & E_{13} \\ E_{21} & A_{22} & E_{23} \\ E_{31} & E_{32} & A_{33} \end{bmatrix},$$

where A_{11} corresponds to eigenvalues inside $(-1, 1)$, A_{22} eigenvalues with absolute value one, and A_{33} eigenvalues outside $[-1, 1]$, then the convergence of $[E_{12}, E_{13}]$ to zero is still cubic.

5. Implementation details and related issues. In this section we address several issues involved in implementing the Algorithm Cubic proposed in section 3. We will also consider how to extend Algorithm Cubic to compute the complete eigendecomposition of an Hermitian matrix and how to handle the generalized eigenvalue problem.

Other intervals. In the case that we want to compute the invariant subspace corresponding to eigenvalues inside the interval (a, b), we can use the following transformation to map (a, b) to $(-1, 1)$:

$$p(x) = \frac{2}{b-a}\left(x - \frac{a+b}{2}\right).$$

It is easy to check that $|p((a, b))| < 1$ and $|p((-\infty, \infty) - [a, b])| > 1$.

Complexity. In step 2 of Algorithm Cubic, each iteration needs five matrix–matrix multiplications of order n matrices and one QR decomposition of a $2n \times n$ matrix. The number of matrix multiplications can be reduced to four if we use the symmetry of the matrices $H_k^H A H_k$ and $G_k^H G_k$.

Storage. We first count the number of storage blocks of size n^2 needed in Algorithm Cubic. The matrix A needs to be stored. We need two blocks for storing $H_k^H A H_k$ and $G_k^H G_k$ which are used for the QR decomposition in step 2.1. We also need two blocks for updating G_{k+1} and H_{k+1} in step 2.2. Therefore, the total storage space needed is $5n^2$. Using the symmetry of $A, H_k^H A H_k$, and $G_k^H G_k$, the storage space can be reduced to $3.5n^2$.

The computation of the QR decomposition in step 2.1 can be done either using a modified Gram–Schmidt method [4, section 5.2.8] or Householder transformations. It is easy to see that the Gram–Schmidt method allows the generation of the orthonormal matrix *in place*, one column at a time. Now we demonstrate that this is also the case with Householder transformations. Let $F \in \mathcal{C}^{m\times n}$ with $m \geq n$. In the QR decomposition, a sequence of Householder vectors $u_1, u_2, \ldots, u_n$ are generated so that

$$H(u_n)H(u_{n-1})\cdots H(u_2)H(u_1)F = [R^H, 0]^H,$$

where $H(u_i) = I_m - 2u_i u_i^H$ and $R \in \mathcal{C}^{n\times n}$ is upper triangular. The matrix R is stored in the upper triangular part of F, and those Householder vectors $u_1, u_2, \ldots, u_n$ are stored in the lower triangular part of F. Now remember all we need is the first n columns of the matrix $H \equiv H(u_1)H(u_2)\cdots H(u_{n-1})H(u_n)$, i.e., $He_i, i = 1, \ldots, n$, where e_i is the ith column of I_m. The key observation here is the fact that the vectors $u_{i+1}, \ldots, u_n$ are not needed in generating He_i, since the first $i-1$ components of u_i are zero. Therefore, we can generate $\{He_i\}$ one column at a time starting with the last column He_n. In LAPACK implementation, a block version is used and the Householder vectors are grouped into block columns of certain size. The above technique can also be modified to handle the block case.

A brief comparison with ISDA. In [1, 3], a simple and elegant method, the so-called invariant subspace decomposition algorithm (ISDA), was proposed to compute invariant subspaces of a symmetric matrix A (some extensions were also considered in [7].) The method first applies a linear transformation to A so that the spectrum of the transformed matrix $\hat{A}$ is within $[0, 1]$. Then a version of truncated incomplete Beta function such as $B_1(x) = 3x^2 - 2x^3$ is repeatedly applied to $\hat{A}$,

$$A_{k+1} = B_1(A_k), \quad k = 0, 1, 2, \ldots, \quad A_0 = \hat{A};$$

the idea is to map the eigenvalues of $\hat{A}$ inside $[0, 1/2)$ to zero and those inside $(1/2, 1]$ to one. In the following we will just point out some differences of the two algorithms. Compared with Algorithm Cubic, ISDA is much simpler and only requires matrix–matrix multiplication in each iteration. The convergence rate of ISDA is quadratic while that of Algorithm Cubic is cubic. However, the convergence mechanism of

the two algorithms are rather different although both of them suffer slow convergence when the eigenvalues are clustered around certain critical points; for ISDA, the critical point is $1/2$, and for Algorithm Cubic, they are -1 and 1. Certain preprocessing steps will alleviate to some extent the slow convergence problem. But these measures will not completely eliminate the problem. The linear transformation step of ISDA might create a tighter cluster since it will squeeze the spectrum of A into $[0, 1]$. It seems that Algorithm Cubic can be readily extended to handle the SVD case, operating only on matrices of order n. It is certainly worthwhile to develop methods that will combine the strength of both algorithms.

The complete eigendecomposition. A complete eigendecomposition of an Hermitian matrix $A \in \mathcal{C}^{n\times n}$ can be obtained by recursively applying Algorithm Cubic. In the following we outline an algorithm which will be used in the numerical examples in section 6. The algorithm starts by applying a linear transformation to A.

Algorithm CubicEig.

1. If $n \le n_{\min}$, compute the eigendecomposition $A = UDU^H$ by a conventional method such as QR algorithm.
2. Transform A to
$$A_0 = \frac{2}{b-a}\left(A - \frac{a+b}{2}I\right),$$
where $a = -\|A\|_F$, and $b = \mathrm{trace}(A)/n$.
3. Compute the invariant subspace of A_0 using Algorithm Cubic to obtain a unitary matrix A and an integer r such that
$$Q^H A Q = \begin{bmatrix} A_{11} & E_{12} \\ E_{21} & A_{22} \end{bmatrix}, \quad A_{11} \in \mathcal{C}^{r\times r}.$$
4. Compute the eigendecomposition of A_{11} and A_{22},
$$A_{11} = U_1 D_1 U_1^H, \quad A_{22} = U_2 D_2 U_2^H,$$
by recursively calling Algorithm CubicEig.
5. Set $U = [Q_1U_1, Q_2U_2]$ and $D = \mathrm{diag}(D_1, D_2)$.

Notice that when n is small enough, i.e., $n \le n_{\min}$, we will switch to a conventional method for computing the eigendecomposition.

Parallel implementation issues. Algorithm CubicEig fits very well with the general structure of divide-and-conquer algorithms. A parallel implementation of it can be modeled after that proposed in [4, section 8.6.5] using a binary computation tree. Here the gluing operations are even simpler: it amounts to modifying the orthogonal matrices at the current level using the eigenvector matrices computed from the next level. Implementation of the primitive matrix operations such as matrix–matrix multiplication and QR decomposition can be found in the package ScaLAPACK [9].

Extension to generalized eigenvalue problem. We consider how to extend Algorithm Cubic to handle the symmetric positive definite generalized eigenvalue problem

$$Ax = \lambda Bx,$$

where A and B are Hermitian, and B is positive definite. We need to modify steps 2.1 and 2.2 as follows:

2.1 Compute the QR decomposition of

$$\begin{bmatrix} -G_k^H B G_k \\ H_k^H A H_k \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} R_k \\ 0 \end{bmatrix}.$$

2.2 Modify G_k and H_k

$$G_{k+1} = G_k Q_{22},$$
$$H_{k+1} = H_k Q_{12}.$$

Let the Cholesky decomposition of B be $B = LL^H$. Upon convergence, we compute the QR decomposition with column pivoting of $L^H G_p = QR\Pi$, and form $V = L^{-H}Q$, then $V^H BV = I$, and $V^H AV$ is block diagonal. Notice that we obtained two smaller ordinary eigenvalue problems.

6. Numerical results. In this section, we use several numerical examples to illustrate the convergence behaviors of Algorithm Cubic. All the computation reported in this section were carried out on a SPARC20 Workstation using MATLAB Version 4.1.

Example 1. In this example, we test randomly generated matrices with eigenvalues well separated from 1 and -1. The matrices are generated as $A = QDQ^T$ with Q orthogonal and D is a diagonal matrix with specified diagonal entries. All the matrices have order $n = 100$, and exactly half of the eigenvalues are inside $(-1, 1)$ and half of them are outside. We first choose D as

$$[1 + \texttt{rand}(1, \texttt{n}/2), -(1 - \texttt{rand}(1, \texttt{n}/2))],$$

where `rand()` gives a uniform distribution on the interval $(0.0, 1.0)$. About 50 matrices from this class were tested, and on average it takes about 10 steps for Algorithm Cubic to deliver $\|E_{21}\|_2/\|A\|_2 \approx O(10^{-15})$. Another class of matrices have D of the form

$$[10 * (1 + \texttt{rand}(1, \texttt{n}/2)), -(1 - \texttt{rand}(1, \texttt{n}/2))/10].$$

About 50 matrices from this class were tested, and on average it takes about three steps for Algorithm Cubic to deliver $\|E_{21}\|_2/\|A\|_2 \approx O(10^{-15})$. We observe that the convergence of the algorithm can be very fast when the eigenvalues of the matrix A are far from the two points 1 and -1.

Example 2. In this example, we illustrate the mixed convergence phenomenon discussed in section 4. The matrix $A = QDQ^T$ is of order $n = 6$, and the matrix D has the form

$$[1.1, 1 - 10 * \texttt{mu}, -1 - \texttt{mu}, -1 + \texttt{tau}, 1 - 2 * \texttt{tau}, 1 - 3 * \texttt{tau}],$$

where `mu` and `tau` are parameters that can be altered. Again, exactly half of the eigenvalues of A are inside $(-1, 1)$ and half of them are outside. At step 2.2, we compute QR decomposition with column pivoting of G_k and H_k as follows: $G_k = Q_k R_k \Pi_k$ and $H_k = \hat{Q}_k \hat{R}_k \hat{\Pi}_k$. Then we compute $A_G^{(k)} = Q_k^H A Q_k$ and $A_H^{(k)} = \hat{Q}_k^H A \hat{Q}_k$. Set

$$\texttt{normAG(k)} = \|A_G^{(k)}(4:6, 1:3)\|_F, \quad \texttt{normAH(k)} = \|A_H^{(k)}(4:6, 1:3)\|_F.$$

The plots of `normAG(k)` and `normAH(k)` are given in Figure 1 for three different choices of `mu` and `tau`.

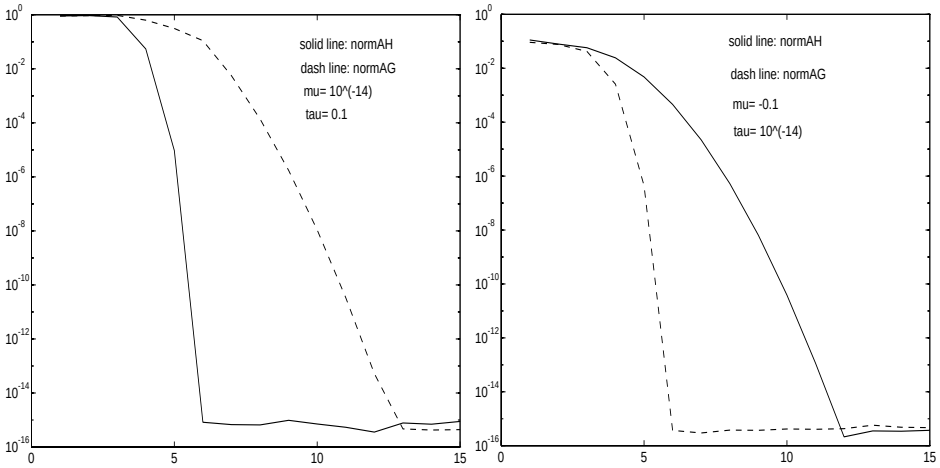


FIG. 1. *Comparison of convergence behavior.*

Example 3. In this example, we demonstrate the possibility of employing a preprocessing technique to accelerate the convergence of Algorithm Cubic. A similar idea has also been used in [2, 3]. The matrix $A = QDQ^T$ is of order $n = 6$, and the matrix D has the form

$$[1.5, 1 + 10^{(-14)}, -1 - 10^{(-15)}, 1 - 10^{(-15)}, -1 + 10^{(-15)}, -0.5].$$

Notice that A has eigenvalues that are very close to 1 and -1. The purpose of preprocessing is to apply a transform to A so that the eigenvalues of the transformed matrix will be separated from 1 and -1. We first construct a polynomial $\phi(x) = (10x^2 - 9)x^2$, then we compute $\hat{A} = \phi(A)$. We apply Algorithm Cubic with A replaced by $\hat{A}$ with only a few steps in the loop step 2, say $k = k_m$. We then continue the loop but with $\hat{A}$ replaced by A. The k_m steps with $\hat{A}$ constitutes the preprocessing phase. The convergence behaviors of Algorithm Cubic with and without preprocessing are plotted in Figure 2.

Example 4. In this example, we apply Algorithm CubicEig to a set of randomly generated matrices and compare its accuracy with that obtained using the MATLAB function `schur`. The matrices are generated as follows:

$$D = \texttt{rand}(\texttt{n}, 1) * 10, \quad [Q_0, R] = \texttt{qr}(\texttt{rand}(\texttt{n})), \quad A = Q_0 \operatorname{diag}(D) Q_0^H.$$

We start with matrices of order $n = 100$ and increment by 10 until $n = 200$; for each matrix size, we test 20 matrices. Let Q and U be the unitary matrices obtained from Algorithm CubicEig with $n_{\min} = 1$ and the MATLAB function `schur`, respectively. We compute the residuals

$$\|Q^H AQ - \operatorname{diag}(Q^H AQ)\|_F, \quad \|U^H AU - \operatorname{diag}(U^H AU)\|_F.$$

Figure 3 plots the residuals for each of the matrix orders. The two sets of residuals are rather comparable with each other.

7. Concluding remarks. We proposed a cubically convergent algorithm, Algorithm Cubic, for computing an invariant subspace of an Hermitian matrix. Once we have block diagonalized a matrix A, we can recursively apply Algorithm Cubic to

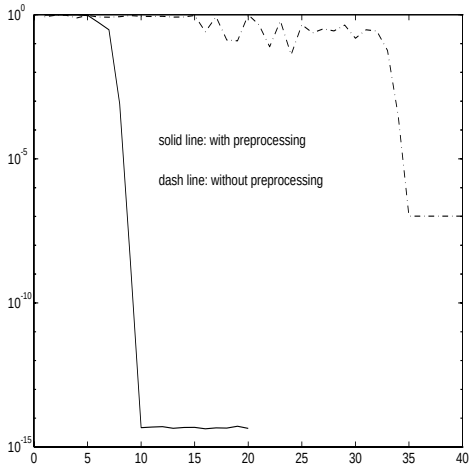


FIG. 2. *The effect of preprocessing.*

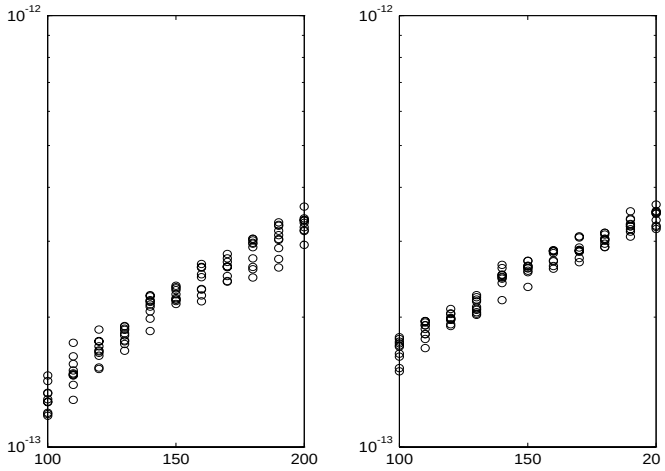


FIG. 3. *(Left) Residuals for Algorithm CubicEig; (Right) Residuals for* MATLAB *function* `schur`.

each of the diagonal blocks. This in effect gives us an algorithm, Algorithm CubicEig, for computing the eigenvalue decomposition of A. A detailed convergence analysis of Algorithm Cubic was given that demonstrates its cubic convergence rate. A general conclusion that can be drawn from the analysis and the numerical experiments is that once the eigenvalues of A are cleared away from 1 and -1, convergence of Algorithm Cubic can occur within just a few steps, around three or four steps at most for matrices of order around a few hundred. If A has eigenvalues very close to either 1 or -1, the convergence of Algorithm Cubic can be painfully slow, and the quality of the block diagonalization deteriorates. One can, however, use a preprocessing step to alleviate to some extent the slow convergence problem. If the goal is to find the eigenvalue decomposition of A, it makes much sense to first divide the spectrum of A into two clusters, and map the interval that contains one of the clusters into the interval $(-1, 1)$. Another interesting phenomenon we observed is that even in the case where both η_- and η_+ are very close to one, which means Algorithm Cubic converges

very slowly, the matrices $Q_k^H A Q_k$ and/or $\hat{Q}_k^H A \hat{Q}_k$ converge quickly to block diagonal form although the sizes of the diagonal blocks will not be r or $n-r$. This points to another possible way to overcome the slow convergence of Algorithm Cubic in those difficult situations: deflate $Q_k^H A Q_k$ and/or $\hat{Q}_k^H A \hat{Q}_k$ when they become block diagonal and work on the diagonal blocks. We need to extend the convergence analysis results in order to better understand the convergence behavior of each individual eigenvalue. Those two topics will be dealt with in a forthcoming paper.

Acknowledgments. The authors want to thank the referees for their careful reading of the paper. Their insightful comments and criticisms greatly improved the presentation of the paper.

REFERENCES

[1] L. Auslander and A. Tsao, *On parallelizable eigensolvers*, Adv. in Appl. Math., 13 (1992), pp. 253–261.

[2] Z. Bai, J. Demmel, and M. Gu, *Inverse Free Parallel Spectral Divide and Conquer Algorithms for Nonsymmetric Eigenproblems*, Tech. report CSD-94-793, Computer Science Division, University of California at Berkeley, Berkeley, CA, 1994.

[3] C. Bischof, S. Huss-Lederman, X. Sun, A. Tsao, and T. Turnbull, *Parallel studies of the invariant subspace decomposition approach for banded symmetric matrices*, in Proc. Seventh SIAM Conference on Parallel Processing for Scientific Computing, San Francisco, CA, 1995, pp. 516–521.

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

[5] N. J. Higham, *The matrix sign decomposition and its relation to the polar decomposition*, Linear Algebra Appl., 212/213 ('994), pp. 3–20.

[6] J. Howland, *The sign matrix and the separation of matrix eigenvalues*, Linear Algebra Appl., 49 (1983), pp. 221–232.

[7] S. Huss-Lederman, A. Tsao, and T. Turnbull, *A parallelizable eigensolver for real diagonalizable matrices with real eigenvalues*, SIAM J. Sci. Comput., 18 (1997), pp. 869–885.

[8] A. Malyshev, *Parallel algorithm for solving some spectral problems of linear algebra*, Linear Algebra Appl., 188/189 (1993), pp. 489–520.

[9] ScaLAPACK, available online from http://www.netlib.org/scalapack/index.html.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 487–498, April 1998

013

SIMULTANEOUS CONTRACTIBILITY*

TSUYOSHI ANDO† AND MAU-HSIANG SHIH‡

Abstract. Let $\mathcal{C}$ be a set of $n \times n$ complex matrices. For $m = 1, 2, \ldots,$ $\mathcal{C}^m$ is the set of all products of matrices in $\mathcal{C}$ of length m. Denote by $\hat{r}(\mathcal{C})$ the joint spectral radius of $\mathcal{C}$, that is,

$$\hat{r}(\mathcal{C}) \stackrel{\text{def}}{=} \limsup_{m\to\infty} [\sup_{A\in\mathcal{C}^m} \|A\|]^{\frac{1}{m}}.$$

We call $\mathcal{C}$ *simultaneously contractible* if there is an invertible matrix S such that

$$\sup\{\|S^{-1}AS\|;\ A \in \mathcal{C}\} < 1,$$

where $\|\cdot\|$ is the spectral norm. This paper is primarily devoted to determining the *optimal joint spectral radius range for simultaneous contractibility* of *bounded sets of $n \times n$ complex matrices,* that is, the maximum subset J of $[0, 1)$ such that if $\mathcal{C}$ is a bounded set of $n \times n$ complex matrices and $\hat{r}(\mathcal{C}) \in J$, then $\mathcal{C}$ is simultaneously contractible. The central result proved in this paper is that this maximum subset is $[0, \frac{1}{\sqrt{n}})$. Our method of proof is based on a matrix-theoretic version of complex John's ellipsoid theorem and the generalized Gelfand spectral radius formula.

Key words. simultaneous contractibility, joint spectral radius, optimal joint spectral radius range, positive definite matrices, Rota's theorem, John's ellipsoid theorem

AMS subject classifications. 15A18, 15A48, 15A60, 34K20, 47A30, 52A20

PII. S0895479897318812

1. Introductory remarks.

Let $\mathbf{C}^n$ be an n-dimensional complex linear space. The inner product

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i \overline{y_i} \quad (x, y \in \mathbf{C}^n)$$

and the associated norm

$$\|x\| = \langle x, x\rangle^{1/2} \quad (x \in \mathbf{C}^n)$$

make $\mathbf{C}^n$ into an n-dimensional Hilbert space, which is denoted by $l_n^2(\mathbf{C})$. For an $n\times n$ complex matrix A, $r(A)$ stands for the spectral radius of A and $\|A\|$ for the spectral norm, the operator norm of A associated with the $l_n^2(\mathbf{C})$ norm $\|x\|$. Let $\mathcal{C}$ be a set of $n \times n$ complex matrices. For $m = 1, 2, \ldots,$ $\mathcal{C}^m$ is the set of all products of matrices in $\mathcal{C}$ of length m. Denote by $\hat{r}(\mathcal{C})$ the *joint spectral radius* of $\mathcal{C}$ [8], that is,

$$\hat{r}(\mathcal{C}) \stackrel{\text{def}}{=} \limsup_{m\to\infty} [\sup_{A\in\mathcal{C}^m} \|A\|]^{\frac{1}{m}}.$$

The quantity $\hat{r}(\mathcal{C})$ *does not* depend on the choice of a norm (since all norms are equivalent on a finite-dimensional space).

*Received by the editors March 19, 1997; accepted for publication (in revised form) by G. P. Styan April 8, 1997.

http://www.siam.org/journals/simax/19-2/31881.html

†Faculty of Economics, Hokusei Gakuen University, Atsubetsu-ku, Sapporo 004-0042, Japan (ando@hokusei.ac.jp). This work was supported in part by Grant-in-Aid for Scientific Research.

‡Department of Mathematics, Chung Yuan Christian University, Chung-Li, Taiwan 32023 (mhshih@poincare.cycu.edu.tw). This work was supported in part by the National Science Council of the Republic of China.

Let us call $\mathcal{C}$ *simultaneously contractible* if there is an invertible matrix S such that

$$\sup\{\|S^{-1}AS\|;\ A \in \mathcal{C}\} < 1.$$

For two complex matrices A and B, the order relation $A \leq B$ (or $B \geq A$) means that $B - A$ is positive semidefinite. The strict inequality $A < B$ (or $B > A$) means that $B - A$ is positive definite. With the ordering "$\leq$", it is readily proved that the simultaneous contractibility of $\mathcal{C}$ is equivalent to the existence of a positive definite matrix H and $0 < \gamma < 1$ such that

$$A^*HA \leq \gamma H \quad (A \in \mathcal{C}).$$

Indeed if we take

$$\sup\{\|S^{-1}AS\|;\ A \in \mathcal{C}\} \equiv \sqrt{\gamma} < 1,$$

then with $H = (SS^*)^{-1}$

$$A^*HA \leq \gamma H \quad (A \in \mathcal{C}).$$

Conversely, if all of these inequalities hold, let $S = H^{-1/2}$.

We now make two general remarks concerning the simultaneous contractibility. First, a simultaneously contractible family $\mathcal{C}$ is necessarily bounded, namely,

$$\|A\| < \|S\| \cdot \|S^{-1}\| \quad (A \in \mathcal{C}).$$

Second, if $\mathcal{C}$ is simultaneously contractible by an invertible matrix S, the multiplicative semigroup generated by $\mathcal{C}$ as well as the convex span of $\mathcal{C}$ are simultaneously contractible by the same matrix S.

Our starting point of this paper is furnished by the classical construction of Rota [7]. If $\mathcal{C}$ consists of a single matrix A, then $\hat{r}(\mathcal{C}) = r(A)$. If $r(A) < 1$, then

$$H \overset{\text{def}}{=} \sum_{k=0}^{\infty} A^{*k}A^k$$

is a well-defined positive definite matrix such that

$$A^*HA < H.$$

This is generalized as follows. If $\mathcal{C}$ consists of m complex matrices, $\mathcal{C} = \{A_1, \ldots, A_m\}$, and $\hat{r}(\mathcal{C}) \leq \frac{\gamma}{\sqrt{m}}$ with $\gamma < 1$, then

$$H \overset{\text{def}}{=} I + \sum_{k=1}^{\infty} \sum_{A \in \mathcal{C}^k} A^*A$$

is a well-defined positive definite matrix such that

$$A_i^*HA_i \leq \gamma^2 H \quad (i = 1, \ldots, m).$$

The constant $\frac{1}{\sqrt{m}}$ is optimal in the sense that there is a set $\mathcal{C}$ consisting of m complex matrices such that $\hat{r}(\mathcal{C}) = \frac{1}{\sqrt{m}}$, but $\mathcal{C}$ is not simultaneously contractible (see the

remark after the proof of Theorem 1.1 in section 3). Incidentally, this optimality also serves to indicate the condition $\hat{r}(\mathcal{C}) < 1$ *does not* necessarily imply the simultaneous contractibility of $\mathcal{C}$ even if $\mathcal{C}$ is multiplicative and convex. If $\hat{r}(\mathcal{C}) \geq 1$, then clearly $\mathcal{C}$ is *not* simultaneously contractible.

The above discussion suggests the following problem. To begin with, the *optimal joint spectral radius range for simultaneous contractibility of bounded sets of $n \times n$ complex matrices* is defined as the maximum subset J of $[0, 1)$ such that if $\mathcal{C}$ is a bounded set of $n \times n$ complex matrices and $\hat{r}(\mathcal{C}) \in J$, then $\mathcal{C}$ is simultaneously contractible. Accordingly, *the optimal joint spectral radius range for simultaneous contractibility of sets of m complex matrices is* $[0, \frac{1}{\sqrt{m}})$. In this note we shall consider the somewhat more complicated question of the determination of the optimal joint spectral radius range for *bounded sets* of $n \times n$ complex matrices. Our principal result is the following.

THEOREM 1.1. *The optimal joint spectral radius range for simultaneous contractibility of bounded sets of $n \times n$ complex matrices is* $[0, \frac{1}{\sqrt{n}})$.

The method of proof employed involves a matrix-theoretic version of complex John's ellipsoid theorem as well as the generalized Gelfand spectral radius formula.

2. A matrix-theoretic version of complex John's ellipsoid theorem. To prove Theorem 1.1, we need a matrix-theoretic version of complex John's ellipsoid theorem. Let us recall that a set E in $l_n^2(\mathbf{C})$ is an *ellipsoid* if there exist a vector $a \in l_n^2(\mathbf{C})$ and a positive definite matrix Q such that

$$E \equiv E(Q, a) \stackrel{\text{def}}{=} \{x \in l_n^2(\mathbf{C});\ \langle Q(x-a), x-a\rangle \leq 1\}.$$

A set K in $l_n^2(\mathbf{C})$ is *balanced with respect to $a \in l_n^2(\mathbf{C})$* if

$$x \in K \Longrightarrow e^{i\theta}(x-a) + a \in K \quad (0 \leq \theta < 2\pi).$$

John's ellipsoid theorem [6]. *If K is a convex body ($\equiv$ compact convex set with nonempty interior) in $l_n^2(\mathbf{R})$ which is (real) balanced with respect to $a \in l_n^2(\mathbf{R})$, then there is a (real) ellipsoid $E(Q, a)$ such that*

$$E(nQ, a) \subset K \subset E(Q, a).$$

John's proof was based on Lagrange's multiplier rule where the subsidiary conditions are inequalities. A complex version of John's ellipsoid theorem is seen in Tomczak-Jaegermann [10, p. 54] in terms of the notion of *Banach–Mazur distance*. We present a matrix-theoretic proof of a complex version of John's theorem. We mention here that John's (real) ellipsoid theorem plays a fundamental role in the study of ellipsoid method for linear programming and of algorithms for convex body (see [5] and [2]).

THEOREM 2.1. *Let $\{A_\lambda;\ \lambda \in \Lambda\}$ be a bounded set of $n \times n$ positive semidefinite matrices such that*

$$\sup_{\lambda \in \Lambda} \langle A_\lambda x, x\rangle > 0 \quad (x \in l_n^2(\mathbf{C}),\ x \neq 0). \tag{1}$$

Then there is an $n \times n$ positive definite matrix H_0 such that

$$\langle H_0 x, x\rangle \geq \sup_{\lambda \in \Lambda} \langle A_\lambda x, x\rangle \geq \frac{1}{n}\langle H_0 x, x\rangle \quad (x \in l_n^2(\mathbf{C})). \tag{2}$$

To prove Theorem 2.1, we need the following construction.

LEMMA 2.2. *Let A be an $n \times n$ complex matrix with $I \geq A \geq 0$ and let $0 \leq \alpha < \frac{1}{n}$. If for some unit vector e_1*

$$\langle Ae_1, e_1 \rangle \leq \alpha,$$

then, with the rank 1 projection $P_1 \stackrel{\text{def}}{=} e_1 \otimes e_1^*$, the matrix

$$T \stackrel{\text{def}}{=} n\alpha P_1 + \frac{n(1-\alpha)}{n-1}(I - P_1)$$

satisfies

$$T \geq A \quad \textit{and} \quad \det(T) < 1.$$

Proof. By the arithmetic-geometric means inequality we have

$$\det(T)^{1/n} = \left\{ n\alpha \times \left(\frac{n(1-\alpha)}{n-1} \right)^{n-1} \right\}^{1/n} < \frac{n\alpha + n(1-\alpha)}{n} = 1.$$

Here strict inequality occurs because the equality implies

$$n\alpha = \frac{n(1-\alpha)}{n-1}, \text{ hence } n\alpha = 1.$$

We have to prove that

$$\langle Tx, x \rangle \geq \langle Ax, x \rangle \quad (x \in l_n^2(\mathbf{C})).$$

To see this, fix x_0. Then there is a unit vector e_2 (depending on x_0), *orthogonal* to e_1, such that x_0 is in the linear hull of e_1 and e_2. Let

$$P_2 \stackrel{\text{def}}{=} e_2 \otimes e_2^*, \text{ and } P \stackrel{\text{def}}{=} P_1 + P_2.$$

Then P is a projection of rank 2. Hence to establish that

$$\langle Tx_0, x_0 \rangle \geq \langle Ax_0, x_0 \rangle,$$

we need to prove only the following inequality:

$$PTP \geq PAP. \tag{3}$$

The matrices PAP and PTP can be considered as 2×2 matrices:

$$PAP = \begin{bmatrix} a & c \\ \bar{c} & b \end{bmatrix}, \quad PTP = \begin{bmatrix} n\alpha & 0 \\ 0 & \frac{n(1-\alpha)}{n-1} \end{bmatrix}.$$

To prove (3) it suffices to show that

$$n\alpha - a \geq 0, \quad \frac{n(1-\alpha)}{n-1} - b \geq 0, \tag{4}$$

and

$$(n\alpha - a) \times \left\{ \frac{n(1-\alpha)}{n-1} - b \right\} \geq |c|^2. \tag{5}$$

To prove (4) and (5), we use the inequality $A \geq 0$ as

$$a \geq 0, \quad b \geq 0, \quad \text{and} \quad ab \geq |c|^2, \tag{6}$$

and the inequality $I \geq A$ as

$$1 - a \geq 0, \quad 1 - b \geq 0, \quad \text{and} \quad (1-a)(1-b) \geq |c|^2. \tag{7}$$

Also, the assumption is used as

$$\alpha \geq a, \quad 1 - n\alpha > 0. \tag{8}$$

Now by (7) and (8)

$$n\alpha - a \geq \alpha - a \geq 0,$$

and

$$\frac{n(1-\alpha)}{n-1} - b = 1 + \frac{1-n\alpha}{n-1} - b \geq 1 - b \geq 0,$$

proving (4).

To see (5) (under (4)), we consider two cases separately.

Case 1. $b \leq 1 - \alpha$. Then by (6)

$$\begin{aligned}(n\alpha - a)\left\{\frac{n(1-\alpha)}{n-1} - b\right\} - |c|^2 &\geq (n-1)\alpha\left\{(1-b) + \frac{1-n\alpha}{n-1}\right\} - ab \\ &\geq (n-1)\alpha\left\{\alpha + \frac{1-n\alpha}{n-1}\right\} - \alpha(1-\alpha) \\ &= \alpha\{(n-1)\alpha + (1-n\alpha) - (1-\alpha)\} = 0,\end{aligned}$$

proving (5).

Case 2. $b > 1 - \alpha$. Then by (7)

$$\begin{aligned}(n\alpha - a)\left\{\frac{n(1-\alpha)}{n-1} - b\right\} - |c|^2 &\geq (n\alpha - a)\left\{(1-b) + \frac{1-n\alpha}{n-1}\right\} - (1-a)(1-b) \\ &\geq \alpha(1-n\alpha) - (1-n\alpha)(1-b) \\ &\geq (1-n\alpha)(\alpha - \alpha) = 0,\end{aligned}$$

proving (5). This completes the proof of (3). □

Proof of Theorem 2.1. By boundedness assumption there is $H \geq 0$ such that

$$H \geq A_\lambda \quad (\lambda \in \Lambda). \tag{9}$$

Claim: *There exists a positive definite matrix H_0 which has a minimum determinant among all H satisfying* (9).

Since by boundedness and assumption (1)

$$x \longmapsto \sqrt{\sup_{\lambda \in \Lambda} \langle A_\lambda x, x\rangle}$$

becomes a norm and all norms on $l_n^2(\mathbf{C})$ are equivalent, there are $\rho > \delta > 0$ such that

$$\rho\langle x, x\rangle \geq \sup_{\lambda \in \Lambda} \langle A_\lambda x, x\rangle \geq \delta\langle x, x\rangle \quad (x \in l_n^2(\mathbf{C})).$$

Let $\mathcal{M}$ be the set of H satisfying (9) and $\det(H) \le \rho^n$. Then it is nonempty because it contains ρI, and the minimum eigenvalue of any H in $\mathcal{M}$ is not smaller than δ. Further, $\mathcal{M}$ is a compact set in the space of $n \times n$ complex matrices. In fact, since for $H > 0$

$$\|H\| = \text{maximum eigenvalue of } H \le \frac{\det(H)}{(\text{minimum eigenvalue of } H)^{n-1}},$$

we can conclude that

$$\|H\| \le \frac{\det(H)}{\delta^{n-1}} \le \frac{\rho^n}{\delta^{n-1}} \quad (H \in \mathcal{M}).$$

Therefore, $\mathcal{M}$ is a bounded set, hence a compact set because its closedness is obvious.

Since the determinant function attains its minimum on each nonempty compact set, we can conclude that there is H_0 which has a minimum determinant among all H satisfying (9).

Now let us write

$$\tilde{A}_\lambda \overset{\text{def}}{=} H_0^{-1/2} A_\lambda H_0^{-1/2} \quad (\lambda \in \Lambda).$$

Then we have

$$I \ge \tilde{A}_\lambda \ge 0 \quad (\lambda \in \Lambda).$$

The requirement of minimum determinant leads to the property

$$T \ge \tilde{A}_\lambda \ (\lambda \in \Lambda) \Rightarrow \ \det(T) \ge 1. \tag{10}$$

The first inequality of (2) is already in the definition of H_0. The second inequality of (2) is equivalent to the following:

$$\sup_{\lambda \in \Lambda} \langle \tilde{A}_\lambda x, x \rangle \ge \frac{1}{n} \langle x, x \rangle \quad (x \in l_n^2(\mathbf{C})). \tag{11}$$

If (11) is not valid, there is a unit vector e_1 such that

$$\alpha \equiv \sup_{\lambda \in \Lambda} \langle \tilde{A}_\lambda e_1, e_1 \rangle < \frac{1}{n}.$$

Construct, according to Lemma 2.2, a matrix T with e_1 and α. (A passing remark: each individual $\tilde{A}_\lambda$ plays no role in this construction.) Therefore, T has the following property:

$$T \ge \tilde{A}_\lambda \quad (\lambda \in \Lambda) \quad \text{and} \quad \det(T) < 1,$$

which contradicts (10). This contradiction shows the validity of (11).

This completes the proof. □

3. Proof of Theorem 1.1 of section 1. In order to prove the optimality of the joint spectral radius range, let us recall the generalized Gelfand spectral radius formula. For a bounded set of $n \times n$ complex matrices $\mathcal{C}$, the generalized Gelfand spectral radius formula asserts that

$$\begin{aligned} r(\mathcal{C}) &\overset{\text{def}}{=} \limsup_{m \to \infty} [\sup_{A \in \mathcal{C}^m} r(A)]^{\frac{1}{m}} \\ &= \hat{r}(\mathcal{C}). \end{aligned}$$

This generalized Gelfand spectral radius formula was conjectured by Daubechies and Lagarias [3] and proved by Berger and Wang [1] using tools from ring theory, and then by Elsner [4] using analytic-geometric tools and by Shih, Wu, and Pang [9] using dynamics method.

We now turn to the proof of Theorem 1.1

(I) Claim: *If $\mathcal{C}$ is a bounded set of $n \times n$ complex matrices and $\hat{r}(\mathcal{C}) < \frac{1}{\sqrt{n}}$, then $\mathcal{C}$ is simultaneously contractible.*

We prove this claim by making use of Theorem 2.1.

Proof. Let

$$\hat{r}(\mathcal{C}) < \alpha < \frac{1}{\sqrt{n}}.$$

Choose a positive integer m such that

$$\sup_{A \in \mathcal{C}^m} \|A\| \leq \alpha^m. \tag{12}$$

Since $\mathcal{C}$ is bounded, we can define a norm $|||\cdot|||$ on $l_n^2(\mathbf{C})$ by setting

$$|||x|||^2 \stackrel{\text{def}}{=} \sup \left\{ \|x\|^2 + \frac{1}{\alpha^2}\|B_1 x\|^2 + \cdots + \frac{1}{\alpha^{2(m-1)}}\|B_{m-1}x\|^2; \right.$$
$$\left. B_j \in \mathcal{C}^j,\ j = 1, \ldots, m-1 \right\}, (x \in l_n^2(\mathbf{C})).$$

By (12) we have for $A \in \mathcal{C}$

$$\begin{aligned} |||Ax|||^2 &= \|Ax\|^2 + \frac{1}{\alpha^2} \sup_{B_1 \in \mathcal{C}^1} \|B_1 Ax\|^2 + \cdots + \frac{1}{\alpha^{2(m-1)}} \sup_{B_{m-1} \in \mathcal{C}^{m-1}} \|B_{m-1}Ax\|^2 \\ &\leq \alpha^2 \left(\frac{1}{\alpha^2} \sup_{B_1 \in \mathcal{C}^1} \|B_1 x\|^2 + \cdots + \frac{1}{\alpha^{2(m-1)}} \sup_{B_{m-1} \in \mathcal{C}^{m-1}} \|B_{m-1}x\|^2 + \|x\|^2 \right) \\ &= \alpha^2 |||x|||^2, \end{aligned}$$

so that

$$|||Ax||| \leq \alpha |||x||| \quad (A \in \mathcal{C};\ x \in l_n^2(\mathbf{C})). \tag{13}$$

We associate to any B_k from $\mathcal{C}^k$ $(k = 1, \ldots, m-1)$ an index λ such that

$$A_\lambda \stackrel{\text{def}}{=} I + \sum_{k=1}^{m-1} \frac{1}{\alpha^{2k}} B_k^* B_k.$$

Then

$$|||x|||^2 = \sup_\lambda \langle A_\lambda x, x \rangle \quad (x \in l_n^2(\mathbf{C})).$$

By Theorem 2.1, there is a positive definite matrix H such that

$$\langle Hx, x \rangle \leq |||x|||^2 \leq n \langle Hx, x \rangle \quad (x \in l_n^2(\mathbf{C})) \tag{14}$$

and by (13) and (14)

$$\langle HAx, Ax \rangle \leq |||Ax|||^2 \leq \alpha^2 n \langle Hx, x \rangle \quad (A \in \mathcal{C};\ x \in l_n^2(\mathbf{C})),$$

hence

$$A^*HA \le \alpha^2 nH \quad (A \in \mathcal{C}).$$

This proves the claim (I). □

(II) Claim: *For any* $\alpha \in [\frac{1}{\sqrt{n}}, 1)$ *there is a set* $\mathcal{C} = \{A_1, \dots, A_n\}$ *of* $n \times n$ *complex matrices such that* $\hat{r}(\mathcal{C}) = \alpha$ *and* $\mathcal{C}$ *is not simultaneously contractible.*

Let $\{e_1, \dots, e_n\}$ be the standard basis for $l_n^2(\mathbf{C})$. Let

$$e \stackrel{\text{def}}{=} e_1 + \cdots + e_n,$$

and

$$E_k \stackrel{\text{def}}{=} e \otimes e_k^* \quad (1 \le k \le n).$$

Let

$$W \stackrel{\text{def}}{=} \operatorname{diag}(\omega, \omega^2, \dots, \omega^n), \text{ where } \omega = e^{2\pi\sqrt{-1}/n} \text{ is an } n\text{th root of unity.}$$

Then

$$E_j W^k = \omega^{jk} E_j \quad (j, k = 1, \dots, n), \tag{15}$$

and

$$E_j E_k = E_k \quad (j, k = 1, \dots, n). \tag{16}$$

Let

$$A_k \stackrel{\text{def}}{=} \alpha W^k E_k \quad (k = 1, \dots, n),$$

and

$$\mathcal{C} = \{A_1, \dots, A_n\}.$$

Claim: $\hat{r}(\mathcal{C}) = \alpha$.

First, by (15)

$$r(W^j E_k) = r(E_k W^j) = r(\omega^{jk} E_k) = r(E_k) = 1 \quad (j, k = 1, \dots, n). \tag{17}$$

Let

$$B_j \stackrel{\text{def}}{=} \alpha W^{k_j} E_{k_j} \in \mathcal{C} \ (j = 1, \dots, m).$$

From (15) and (16) we have

$$B_1 B_2 \cdots B_m = \alpha^m \omega^{k_1 k_2 + k_2 k_3 + \cdots + k_{m-1} k_m} W^{k_1} E_{k_m},$$

and so

$$\begin{aligned} r(B_1 B_2 \cdots B_m) &= \alpha^m r(W^{k_1} E_{k_m}) \\ &= \alpha^m \quad \text{by (17).} \end{aligned}$$

Therefore,

$$r(B) = \alpha^m \quad (B \in \mathcal{C}^m;\ m = 1, 2, \dots),$$

and so

$$r(\mathcal{C}) = \alpha.$$

By the generalized Gelfand spectral radius formula, we conclude that

$$\hat{r}(\mathcal{C}) = \alpha.$$

Claim: *$\mathcal{C}$ is not simultaneously contractible.*

Suppose, by contradiction, that there is $H > 0$ such that

$$A_k^* H A_k < H \quad (k = 1, \ldots, n). \tag{18}$$

Denote the entries of H by h_{ij} $(i, j = 1, \ldots, n)$. Then by (18) we have

$$\begin{aligned} 0 < \langle (H - A_k^* H A_k) e_k, e_k \rangle &= \langle H e_k, e_k \rangle - \langle H A_k e_k, A_k e_k \rangle \\ &= h_{kk} - \alpha^2 \langle H W^k e, W^k e \rangle \\ &= h_{kk} - \alpha^2 \sum_{r=1}^{n} \sum_{s=1}^{n} h_{rs} \omega^{(s-r)k} \quad (k = 1, \ldots, n). \end{aligned} \tag{19}$$

Since $\omega = e^{2\pi\sqrt{-1}/n}$, we have

$$\sum_{k=1}^{n} \omega^{jk} = \begin{cases} 0 & \text{if } n \text{ does not divide } j, \\ n & \text{if } n \text{ divides } j. \end{cases} \tag{20}$$

Since $\frac{1}{\sqrt{n}} \leq \alpha < 1$, (19) and (20) together imply that

$$\begin{aligned} 0 &< \sum_{k=1}^{n} h_{kk} - \alpha^2 \sum_{k=1}^{n} \sum_{r=1}^{n} \sum_{s=1}^{n} h_{rs} \omega^{(s-r)k} \\ &= \operatorname{tr}(H) - n\alpha^2 \operatorname{tr}(H) \\ &\leq 0, \end{aligned}$$

in contradiction. This contradiction proves that $\mathcal{C}$ is *not* simultaneously contractible.

This proves the theorem.

Remark. The optimality example shows also that $\frac{1}{\sqrt{m}}$ is optimal in the generalization of Rota's construction mentioned in section 1.

4. Zero joint spectral radius and triangularization. The case of $\hat{r}(\mathcal{C}) = 0$ shows the interesting fact that $\mathcal{C}$ is simultaneously upper triangularizable by a unitary matrix, and so $\mathcal{C}$ is simultaneously contractible if $\mathcal{C}$ is bounded. Notice that $\hat{r}(\mathcal{C}) = 0$ implies simultaneous contractibility of $\mathcal{C}$ by Theorem 1.1.

LEMMA 4.1. *Let $\mathcal{C}$ be a set of $n \times n$ complex matrices. If $\hat{r}(\mathcal{C}) = 0$, then $\mathcal{C}$ is simultaneously upper triangularizable by a unitary matrix, that is, there is a unitary matrix U such that $U^* A U$ is an upper triangular matrix for all A in $\mathcal{C}$.*

Proof. We prove the assertion by induction on the dimension n. The case of $n = 1$ is trivial. Assume that $n > 1$ and the assertion is true for all cases of dimension less than n.

Let $\mathcal{A}$ be the algebra spanned by $\mathcal{C}$.

(i) Claim: $r(A) = 0 \quad (A \in \mathcal{A})$.

Since $\mathcal{A}$ is of finite dimension, there are linearly independent

$$B_i \in C^{k_i} \quad (i = 1, \ldots, N)$$

which spans $\mathcal{A}$. Let

$$p \equiv \min_{1\le i\le N} k_i \quad \text{and} \quad q \equiv \max_{1\le i\le N} k_i.$$

Each $A \in \mathcal{A}$ has a unique representation

$$A = \sum_{i=1}^{N} \alpha_i B_i.$$

Since

$$A^m = \sum_{1\le j_i\le N} \left(\prod_{i=1}^{m} \alpha_{j_i}\right)(B_{j_1}\cdots B_{j_m}),$$

with $\alpha \equiv \max_{1\le i\le N} |\alpha_i|$ we have

$$\|A^m\|^{1/m} \le \alpha N \cdot \max_{mp\le k\le mq} \{\sup_{B\in\mathcal{C}^k} \|B\|^{1/m}\}.$$

Assumption $\hat{r}(\mathcal{C}) = 0$ implies that for any $0 < \epsilon < 1$ there is m_0 such that

$$\sup_{B\in\mathcal{C}^k} \|B\|^{1/k} \le \epsilon \quad (k \ge m_0).$$

Then we have for $k \ge mp$ and $m \ge m_0$

$$\sup_{B\in\mathcal{C}^k} \|B\|^{1/m} \le \epsilon^{k/m} \le \epsilon^p$$

so that

$$\|A^m\|^{1/m} \le \alpha N \epsilon^p \quad (m \ge m_0).$$

Since $0 < \epsilon < 1$ is arbitrary, this implies

$$\|A^m\|^{1/m} \to 0 \ as \ m \to \infty.$$

(ii) Claim: *There is a nontrivial subspace $M \subset l_n^2(\mathbf{C})$ which is invariant for all $A \in \mathcal{A}$.*

This claim is trivial if $\mathcal{A} = \{0\}$. If $\mathcal{A} \ne \{0\}$, there is $x_0 \in l_n^2(\mathbf{C})$ such that $M \overset{\text{def}}{=} \mathcal{A}x_0 \ne \{0\}$. Since $\mathcal{C}\mathcal{A} \subset \mathcal{A}$, this subspace M is invariant for all $A \in \mathcal{A}$. The subspace M does not coincide with the whole space $l_n^2(\mathbf{C})$. For otherwise $x_0 \in M$, that is, there is $A \in \mathcal{A}$ such that $x_0 = Ax_0$, in contradiction to $r(A) = 0$, guaranteed in claim (i). Thus, M meets the requirement.

(iii) Since M in (ii) is invariant for all $A \in \mathcal{C}$, according to the orthogonal decomposition

$$l_n^2(\mathbf{C}) = M \oplus M^{\perp},$$

$A \in \mathcal{A}$ is represented in the block matrix form

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

that is, there is a unitary matrix V, common for all $A \in \mathcal{C}$, such that

$$V^*AV = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

(iv) Let $0 < \dim(M) \equiv m < n$. Since in the representation in (iii)

$$\mathcal{C}_1 \equiv \{A_{11};\ A \in \mathcal{C}\},\ \mathcal{C}_2 \equiv \{A_{22};\ A \in \mathcal{C}\}$$

are sets of complex matrices of sizes m and $n-m$, respectively, and

$$\hat{r}(\mathcal{C}_i) \le \hat{r}(\mathcal{C}) = 0 \quad (i = 1, 2),$$

by induction assumption there are unitary matrices V_1 of order m (respectively, V_2 of order $n-m$) such that $V_1^*A_{11}V_1$ (respectively, $V_2^*A_{22}V_2$) is an upper triangular matrix of order m (respectively, $n-m$) for all $A_{11} \in \mathcal{C}_1$ (respectively, all $A_{22} \in \mathcal{C}_2$). Let

$$U \equiv V \cdot (V_1 \oplus V_2).$$

Then we can conclude from the above that U^*AU is an upper triangular matrix for all $A \in \mathcal{C}$.

This completes the proof. □

LEMMA 4.2. *Let $\mathcal{C}$ be a bounded set of $n \times n$ complex matrices. If $\mathcal{C}$ is simultaneously upper triangularizable, then for each $\epsilon > 0$ there is an invertible matrix P such that*

$$\sup\{\|P^{-1}AP\|;\ A \in \mathcal{C}\} \le \sup\{r(A);\ A \in \mathcal{C}\} + \epsilon.$$

Proof. Let Q be an invertible matrix such that

$$Q^{-1}AQ = [a_{ij}]$$

is upper triangular for all A in $\mathcal{C}$ and a_{ii} $(1 \le i \le n)$ are eigenvalues of A.

Let $\epsilon > 0$ be given. For $0 < \delta < \epsilon$, let

$$D \overset{\text{def}}{=} \operatorname{diag}(1, \delta, \ldots, \delta^{n-1}).$$

Then

$$D^{-1}Q^{-1}AQD = \begin{bmatrix} a_{11} & \delta a_{12} & & \cdots & \delta^{n-1}a_{1n} \\ & a_{22} & \delta a_{23} & \cdots & \delta^{n-2}a_{2n} \\ & & \ddots & \vdots & \vdots \\ & \mathrm{O} & & & \delta a_{n-1,n} \\ & & & & a_{nn} \end{bmatrix}.$$

Remark that for any matrix $T = [t_{ij}]$

$$\begin{aligned} \|T\| &\le \|\operatorname{diag}(T)\| + \|T - \operatorname{diag}(T)\| \\ &\le \max_{i=1,\ldots,n} |t_{ii}| + \sqrt{\sum_{i \ne j} |t_{ij}|^2}. \end{aligned}$$

Since $\mathcal{C}$ is bounded, for all $A \in \mathcal{C}$ we have

$$\sup_{A \in \mathcal{C}} \|D^{-1}Q^{-1}AQD\| \le \sup\{r(A);\ A \in \mathcal{C}\} + \epsilon \text{ if } \delta > 0 \text{ is small enough.}$$

Let $S \stackrel{\text{def}}{=} QD$. Then

$$\sup\{\|S^{-1}AS\|;\ A \in \mathcal{C}\} \le \sup\{r(A);\ A \in \mathcal{C}\} + \epsilon,$$

completing the proof. □

Combining Lemmas 4.1 and 4.2, we have the following.

THEOREM 4.3. *Let $\mathcal{C}$ be a bounded set of $n \times n$ complex matrices. If $\hat{r}(\mathcal{C}) = 0$, then for each $\epsilon > 0$ there is an invertible matrix S such that $S^{-1}AS$ $(A \in \mathcal{C})$ are upper triangular and*

$$\sup\{\|S^{-1}AS\|;\ A \in \mathcal{C}\} < \epsilon.$$

COROLLARY 4.4. *If $\mathcal{C}$ is a bounded multiplicative semigroup of $n \times n$ nilpotent complex matrices, then for each $\epsilon > 0$ there is an invertible matrix S such that*

$$\sup\{\|S^{-1}AS\|;\ A \in \mathcal{C}\} < \epsilon.$$

Proof. Since each A in $\mathcal{C}$ is nilpotent, $r(A) = 0$. Since $\mathcal{C}$ is multiplicative,

$$r(\mathcal{C}) = \limsup_{m\to\infty}[\sup_{A\in\mathcal{C}^m} r(A)]^{\frac{1}{m}} = 0.$$

By the generalized Gelfand spectral radius formula, we have $\hat{r}(\mathcal{C}) = 0$. Applying Theorem 4.3, the assertion is proved. □

A set of *commuting* matrices is another example of a simultaneously upper triangularizable set. A proof is easy by induction on the dimension n. For a commuting set $\mathcal{C}$, $\hat{r}(\mathcal{C}) = \sup\{r(A);\ A \in \mathcal{C}\}$ by the generalized Gelfand spectral radius formula.

Now the following theorem is immediate from Lemma 4.2.

THEOREM 4.5. *The optimal joint spectral radius range for simultaneous contractibility of bounded commuting sets of $n \times n$ complex matrices is $[0, 1)$.*

REFERENCES

[1] M. A. BERGER AND Y. WANG, *Bounded semigroups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.

[2] R. G. BLAND, D. GOLDFARB, AND M. J. TODD, *The ellipsoid method: A survey*, Oper. Res., 29 (1981), pp. 1039–1091.

[3] I. DAUBECHIES AND J. C. LAGARIAS, *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.

[4] L. ELSNER, *The generalized spectral radius theorem: An analytic-geometric proof*, Linear Algebra Appl., 220 (1995), pp. 151–159.

[5] M. GRÖTSCHEL, L. LOVÁSZ, AND A. SCHRIJVER, *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, Berlin, Heidelberg, New York, 1988.

[6] F. JOHN, *Extremum problems with inequalities as subsidiary conditions*, in Studies and Essays, Interscience, New York, 1948, pp. 184–204; reprinted in Collected Papers, Vol. 2, J. Moser, ed., Birkhäuser, Boston, MA, 1985, pp. 543–560.

[7] G.-C. ROTA, *On models for linear operators*, Comm. Pure Appl. Math., 8 (1960), pp. 469–472.

[8] G.-C. ROTA AND G. STRANG, *A note on the joint spectral radius*, Indag. Math., 22 (1960), pp. 379–381.

[9] M.-H. SHIH, J.-W. WU, AND C.-T. PANG, *Asymptotic stability and generalized Gelfand spectral radius formula*, Linear Algebra Appl., 252 (1997), pp. 61–70.

[10] N. TOMCZAK-JAEGERMANN, *Banach-Mazur Distances and Finite-Dimensional Operator Ideals,* Pitman Monographs Surveys Pure Appl. Math., 38, John Wiley, New York, 1989.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 499–533, April 1998

014

DETERMINANT MAXIMIZATION WITH LINEAR MATRIX INEQUALITY CONSTRAINTS*

LIEVEN VANDENBERGHE†, STEPHEN BOYD‡, AND SHAO-PO WU‡

Abstract. The problem of maximizing the determinant of a matrix subject to linear matrix inequalities (LMIs) arises in many fields, including computational geometry, statistics, system identification, experiment design, and information and communication theory. It can also be considered as a generalization of the semidefinite programming problem.

We give an overview of the applications of the determinant maximization problem, pointing out simple cases where specialized algorithms or analytical solutions are known. We then describe an interior-point method, with a simplified analysis of the worst-case complexity and numerical results that indicate that the method is very efficient, both in theory and in practice. Compared to existing specialized algorithms (where they are available), the interior-point method will generally be slower; the advantage is that it handles a much wider variety of problems.

Key words. semidefinite programming, interior-point methods, linear matrix inequalities

AMS subject classifications. 65K05, 49M45, 90C25, 90C90, 15A15

PII. S0895479896303430

1. Introduction.

We consider the optimization problem

$$\begin{array}{ll} \text{minimize} & c^T x + \log\det G(x)^{-1} \\ \text{subject to} & G(x) \succ 0 \\ & F(x) \succeq 0, \end{array} \tag{1.1}$$

where the optimization variable is the vector $x \in \mathbf{R}^m$. The functions $G : \mathbf{R}^m \to \mathbf{R}^{l\times l}$ and $F : \mathbf{R}^m \to \mathbf{R}^{n\times n}$ are affine:

$$\begin{aligned} G(x) &= G_0 + x_1 G_1 + \cdots + x_m G_m, \\ F(x) &= F_0 + x_1 F_1 + \cdots + x_m F_m, \end{aligned}$$

where $G_i = G_i^T$ and $F_i = F_i^T$. The inequality signs in (1.1) denote matrix inequalities, i.e., $G(x) \succ 0$ means $z^T G(x) z > 0$ for all nonzero z and $F(x) \succeq 0$ means $z^T F(x) z \geq 0$ for all z. We call $G(x) \succ 0$ and $F(x) \succeq 0$ (strict and nonstrict, respectively) *linear matrix inequalities* (LMIs) in the variable x. We will refer to problem (1.1) as a max-det problem, since in many cases the term $c^T x$ is absent, so the problem reduces to maximizing the determinant of $G(x)$ subject to LMI constraints.

The max-det problem is a convex optimization problem, i.e., the objective function $c^T x + \log\det G(x)^{-1}$ is convex (on $\{x \mid G(x) \succ 0\}$), and the constraint set is convex. Indeed, LMI constraints can represent many common convex constraints, including linear inequalities, convex quadratic inequalities, and matrix norm and eigen-

*Received by the editors April 25, 1996; accepted for publication (in revised form) by M. Overton April 10, 1997. This research was supported in part by AFOSR under F49620-95-1-0318, NSF under ECS-9222391 and EEC-9420565, MURI under F49620-95-1-0525, and NATO Collaborative Research grant CRG-941269. Associated software is available online from http://www-isl.stanford.edu/people/boyd and via anonymous ftp from isl.stanford.edu in pub/boyd/maxdet.

http://www.siam.org/journals/simax/19-2/30343.html

†Electrical Engineering Department, University of California, Los Angeles CA 90095-1594 (vandenbe@ee.ucla.edu).

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

value constraints (see Alizadeh [1], Boyd, et al. [13], Lewis and Overton [47], Nesterov and Nemirovsky [51, sect. 6.4], and Vandenberghe and Boyd [69]).

In this paper we describe an interior-point method that solves the max-det problem very efficiently, both in worst-case complexity theory and in practice. The method we describe shares many features of interior-point methods for linear and semidefinite programming. In particular, our computational experience (which is limited to problems of moderate size — several hundred variables, with matrices up to 100×100) indicates that the method we describe solves the max-det problem (1.1) in a number of iterations that hardly varies with problem size, and typically ranges between 5 and 50; each iteration involves solving a system of linear equations.

Max-det problems arise in many fields, including computational geometry, statistics, and information and communication theory, so the duality theory and algorithms we develop have wide application. In some of these applications, and for very simple forms of the problem, the max-det problems can be solved by specialized algorithms or, in some cases, analytically. Our interior-point algorithm will generally be *slower* than the specialized algorithms (when the specialized algorithms can be used). The *advantage* of our approach is that it is much more general; it handles a much wider variety of problems. The analytical solutions or specialized algorithms, for example, cannot handle the addition of (convex) constraints; our algorithm for general max-det problems does.

In the remainder of section 1, we describe some interesting special cases of the max-det problem, such as semidefinite programming and analytic centering. In section 2 we describe examples and applications of max-det problems, pointing out analytical solutions where they are known, and interesting extensions that can be handled as general max-det problems. In section 3 we describe a duality theory for max-det problems, pointing out connections to semidefinite programming duality. Our interior-point method for solving the max-det problem (1.1) is developed in sections 4–9. We describe two variations: a simple "short-step" method, for which we can prove polynomial worst-case complexity, and a "long-step" or adaptive step predictor-corrector method which has the same worst-case complexity but is much more efficient in practice. We finish with some numerical experiments. For the sake of brevity, we omit most proofs and some important numerical details, and refer the interested reader to the technical report [70]. A C implementation of the method described in this paper is also available [76].

Let us now describe some special cases of the max-det problem.

Semidefinite programming. When $G(x) = 1$, the max-det problem reduces to

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & F(x) \succeq 0, \end{array} \tag{1.2}$$

which is known as a *semidefinite program* (SDP). Semidefinite programming unifies a wide variety of convex optimization problems, e.g., linear programming,

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax \leq b \end{array}$$

which can be expressed as an SDP with $F(x) = \mathbf{diag}(b - Ax)$. For surveys of the theory and applications of semidefinite programming, see [1], [13], [46], [47], [51, sect. 6.4], and [69].

Analytic centering. When $c = 0$ and $F(x) = 1$, the max-det problem (1.1) reduces to

$$\begin{array}{ll} \text{minimize} & \log\det G(x)^{-1} \\ \text{subject to} & G(x) \succ 0, \end{array} \tag{1.3}$$

which we call the *analytic centering* problem. We will assume that the feasible set $\{x \mid G(x) \succ 0\}$ is nonempty and bounded, which implies that the matrices G_i, $i = 1, \ldots, m$, are linearly independent, and that the objective $\phi(x) = \log\det G(x)^{-1}$ is strictly convex (see, e.g., [69] or [12]). Since the objective function grows without bound as x approaches the boundary of the feasible set, there is a unique solution $x^\star$ of (1.3). We call $x^\star$ the *analytic center* of the LMI $G(x) \succ 0$. The analytic center of an LMI generalizes the analytic center of a set of linear inequalities, introduced by Sonnevend [64, 65].

Since the constraint cannot be active at the analytic center, $x^\star$ is characterized by the optimality condition $\nabla\phi(x^\star) = 0$:

$$(\nabla\phi(x^\star))_i = -\mathbf{Tr} G_i G(x^\star)^{-1} = 0, \quad i = 1, \ldots, m \tag{1.4}$$

(see, for example, Boyd and El Ghaoui [12]).

The analytic center of an LMI is important for several reasons. We will see in section 5 that the analytic center can be computed very efficiently, so it can be used as an easily computed robust solution of the LMI. Analytic centering also plays an important role in interior-point methods for solving the more general max-det problem (1.1). Roughly speaking, the interior-point methods solve the general problem by solving a sequence of analytic centering problems.

Parameterization of LMI feasible set. Let us restore the term $c^T x$:

$$\begin{array}{ll} \text{minimize} & c^T x + \log\det G(x)^{-1} \\ \text{subject to} & G(x) \succ 0, \end{array} \tag{1.5}$$

retaining our assumption that the feasible set $\mathbf{X} = \{x \mid G(x) \succ 0\}$ is nonempty and bounded, so the matrices G_i are linearly independent and the objective function is strictly convex. Thus, problem (1.5) has a unique solution $x^\star(c)$, which satisfies the optimality conditions $c + \nabla\phi(x^\star(c)) = 0$, i.e.,

$$\mathbf{Tr} G_i G(x^\star(c))^{-1} = c_i, \quad i = 1, \ldots, m.$$

Thus, for each $c \in \mathbf{R}^m$, we have a (readily computed) point $x^\star(c)$ in the set $\mathbf{X}$.

Conversely, given a point $x \in \mathbf{X}$, define $c \in \mathbf{R}^m$ by $c_i = \mathbf{Tr} G(x)^{-1} G_i$, $i = 1, \ldots, m$. Evidently we have $x = x^\star(c)$. In other words, there is a one-to-one correspondence between vectors $c \in \mathbf{R}^m$ and feasible vectors $x \in \mathbf{X}$: the mapping $c \mapsto x^\star(c)$ is a parameterization of the feasible set $\mathbf{X}$ of the strict LMI $G(x) \succ 0$, with parameter $c \in \mathbf{R}^m$. This parameterization of the set $\mathbf{X}$ is related to the *Legendre transform* of the convex function $\log\det G(x)^{-1}$, defined by

$$\mathcal{L}(y) = -\inf\{-y^T x + \log\det G(x)^{-1} \mid G(x) \succ 0\}.$$

Maximal lower bounds in the positive definite cone. Here we consider a simple example of the max-det problem. Let $A_i = A_i^T$, $i = 1, \ldots, L$, be positive definite matrices in $\mathbf{R}^{p \times p}$. A matrix X is a lower bound of the matrices A_i if $X \preceq A_i$, $i = 1, \ldots, L$; it is a maximal lower bound if there is no lower bound Y with $Y \neq X$, $Y \succeq X$.

Since the function $\log\det X^{-1}$ is monotone decreasing with respect to the positive semidefinite cone, i.e.,

$$0 \prec X \preceq Y \Longrightarrow \log\det Y^{-1} \leq \log\det X^{-1},$$

we can compute a maximal lower bound $A_{\rm mlb}$ by solving

$$\begin{array}{ll} \text{minimize} & \log\det X^{-1} \\ \text{subject to} & X \succ 0 \\ & X \preceq A_i, \quad i = 1, \ldots, L. \end{array} \tag{1.6}$$

This is a max-det problem with $p(p+1)/2$ variables (the elements of the matrix X) and L LMI constraints $A_i - X \succeq 0$, which we can also consider as diagonal blocks of one single block diagonal LMI

$$\mathbf{diag}(A_1 - X, A_2 - X, \ldots, A_L - X) \succeq 0.$$

Of course there are other maximal lower bounds; replacing $\log\det X^{-1}$ by any other monotone decreasing matrix function, e.g., $-\mathbf{Tr}X$ or $\mathbf{Tr}X^{-1}$, will also yield (other) maximal lower bounds. The maximal lower bound $A_{\rm mlb}$ obtained by solving (1.6), however, has the property that it is invariant under congruence transformations, i.e., if the matrices A_i are transformed to TA_iT^T, where $T \in \mathbf{R}^{p\times p}$ is nonsingular, then the maximal lower bound obtained from (1.6) is $TA_{\rm mlb}T^T$.

2. Examples and applications. In this section we catalog examples and applications. The reader interested only in duality theory and solution methods for the max-det problem can skip directly to section 3.

2.1. Minimum volume ellipsoid containing given points. Perhaps the earliest and best-known application of the max-det problem arises in the problem of determining the minimum volume ellipsoid that contains given points $x^1, \ldots, x^K$ in $\mathbf{R}^n$ (or, equivalently, their convex hull $\mathbf{Co}\{x^1, \ldots, x^K\}$). This problem has applications in cluster analysis (Rosen [58], Barnes [9]) and robust statistics (in ellipsoidal peeling methods for outlier detection; see Rousseeuw and Leroy [59, sect. 7]).

We describe the ellipsoid as $\mathcal{E} = \{x \mid \|Ax + b\| \leq 1\}$, where $A = A^T \succ 0$, so the volume of $\mathcal{E}$ is proportional to $\det A^{-1}$. Hence the minimum volume ellipsoid that contains the points x^i can be computed by solving the convex problem

$$\begin{array}{ll} \text{minimize} & \log\det A^{-1} \\ \text{subject to} & \|Ax^i + b\| \leq 1, \ i = 1, \ldots, K \\ & A = A^T \succ 0, \end{array} \tag{2.1}$$

where the variables are $A = A^T \in \mathbf{R}^{n\times n}$ and $b \in \mathbf{R}^n$. The norm constraints $\|Ax^i + b\| \leq 1$, which are just convex quadratic inequalities in the variables A and b, can be expressed as LMIs

$$\left[\begin{array}{cc} I & Ax^i + b \\ (Ax^i + b)^T & 1 \end{array}\right] \succeq 0.$$

These LMIs can in turn be expressed as one large block diagonal LMI, so (2.1) is a max-det problem in the variables A and b.

Nesterov and Nemirovsky [51, sect. 6.5] and Khachiyan and Todd [43] describe interior-point algorithms for computing the maximum volume ellipsoid in a polyhedron described by linear inequalities (as well as the minimum volume ellipsoid covering a polytope described by its vertices).

Many other geometrical problems involving ellipsoidal approximations can be formulated as max-det problems. References [13, sect. 3.7], [16], and [68] give several examples, including the maximum volume ellipsoid contained in the intersection or in the sum of given ellipsoids, and the minimum volume ellipsoid containing the sum of given ellipsoids. For other ellipsoidal approximation problems, suboptimal solutions can be computed via max-det problems.

Ellipsoidal approximations of convex sets are used in control theory and signal processing in *bounded-noise* or *set-membership* techniques. These techniques were first introduced for state estimation (see, e.g., Schweppe [62], [63], Witsenhausen [75], Bertsekas and Rhodes [11], Chernousko [16], [17]) and later applied to system identification (Fogel [35], Fogel and Huang [36], Norton [52], [53, sect. 8.6], Walter and Piet-Lahanier [71], Cheung, Yurkovich, and Passino [18]) and signal processing Deller [23]. (For a survey emphasizing signal processing applications, see Deller, Nayeri, and Odeh [24]).

Other applications include the *method of inscribed ellipsoids* developed by Tarasov, Khachiyan, and Erlikh [66] and design centering (Sapatnekar [60]).

2.2. Matrix completion problems.

Positive definite matrix completion. In a positive definite matrix completion problem we are given a symmetric matrix $A_f \in \mathbf{R}^{n\times n}$, some entries of which are fixed; the remaining entries are to be chosen so that the resulting matrix is positive definite.

Let the positions of the free (unspecified) entries be given by the index pairs (i_k, j_k), (j_k, i_k), $k = 1, \ldots, m$. We can assume that the diagonal elements are fixed, i.e., $i_k \neq j_k$ for all k. (If a diagonal element, say the (l, l)th, is free, we take it to be very large, which makes the lth row and column of A_f irrelevant.) The positive definite completion problem can be cast as an SDP feasibility problem:

$$\begin{array}{ll} \text{find} & x \in \mathbf{R}^m \\ \text{such that} & A(x) \stackrel{\Delta}{=} A_f + \displaystyle\sum_{k=1}^{m} x_k \left(E_{i_k j_k} + E_{j_k i_k}\right) \succ 0, \end{array}$$

where E_{ij} denotes the matrix with all elements zero except the (i, j) element, which is equal to one. Note that the set $\{x \mid A(x) \succ 0\}$ is bounded since the diagonal elements of $A(x)$ are fixed.

Maximum entropy completion. The analytic center of the LMI $A(x) \succ 0$ is sometimes called the *maximum entropy completion* of A_f. From the optimality conditions (1.4), we see that the maximum entropy completion $x^\star$ satisfies

$$2\mathbf{Tr} E_{i_k j_k} A(x^\star)^{-1} = 2\left(A(x^\star)^{-1}\right)_{i_k j_k} = 0, \qquad k = 1, \ldots, m,$$

i.e., the matrix $A(x^*)^{-1}$ has a zero entry in every location corresponding to an unspecified entry in the original matrix. This is a very useful property in many applications; see, for example, Dempster [27] or Dewilde and Ning [30].

Parameterization of all positive definite completions. As an extension of the maximum entropy completion problem, consider

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr} C A(x) + \log\det A(x)^{-1} \\ \text{subject to} & A(x) \succ 0, \end{array} \tag{2.2}$$

where $C = C^T$ is given. This problem is of the form (1.5); the optimality conditions are

$$A(x^*) \succ 0, \quad \left(A(x^*)^{-1}\right)_{i_k j_k} = C_{i_k j_k}, \quad k = 1, \ldots, m, \tag{2.3}$$

i.e., the inverse of the optimal completion matches the given matrix C in every free entry. Indeed, this gives a parameterization of all positive definite completions: a positive definite completion $A(x)$ is uniquely characterized by specifying the elements of its inverse in the free locations, i.e., $\left(A(x)^{-1}\right)_{i_k j_k}$. Problem (2.2) has been studied by Bakonyi and Woerdeman [8].

Contractive completion. A related problem is the contractive completion problem: given a (possibly nonsymmetric) matrix A_f and m index pairs (i_k, j_k), $k = 1, \ldots, m$, find a matrix

$$A(x) = A_f + \sum_{k=1}^{m} x_k E_{i_k, j_k}$$

with spectral norm (maximum singular value) less than one.

This can be cast as a semidefinite programming feasibility problem [69]: find x such that

$$\begin{bmatrix} I & A(x) \\ A(x)^T & I \end{bmatrix} \succ 0. \tag{2.4}$$

One can define a maximum entropy solution as the solution that maximizes the determinant of (2.4), i.e., solves the max-det problem

$$\begin{array}{ll} \text{maximize} & \log\det(I - A(x)^T A(x)) \\ \text{subject to} & \begin{bmatrix} I & A(x) \\ A(x)^T & I \end{bmatrix} \succ 0. \end{array} \tag{2.5}$$

See Naevdal and Woerdeman [50], Helton and Woerdeman [38]. For a statistical interpretation of (2.5), see section 2.3.

Specialized algorithms and references. Very efficient algorithms have been developed for certain specialized types of completion problems. A well-known example is the maximum entropy completion of a positive definite banded Toeplitz matrix (Dym and Gohberg [31], Dewilde and Deprettere [29]). Davis, Kahan, and Weinberger [22] discuss an analytic solution for a contractive completion problem with a special (block matrix) form. The methods discussed in this paper solve the *general* problem efficiently, although they are slower than the specialized algorithms where they are applicable. Moreover, they have the advantage that other convex constraints, e.g., upper and lower bounds on certain entries, are readily incorporated.

Completion problems, and specialized algorithms for computing completions, have been discussed by many authors; see, e.g., Dym and Gohberg [31], Grone, Johnson, Sá and Wolkowicz [37], Barrett, Johnson and Lundquist [10], Lundquist and Johnson [49], Dewilde and Deprettere [29], Dembo, Mallows, and Shepp [26]. Johnson gives a survey in [41]. An interior-point method for an approximate completion problem is discussed in Johnson, Kroschel, and Wolkowicz [42].

We refer to Boyd et al. [13, sect. 3.5] and El Ghaoui [32] for further discussion and additional references.

2.3. Risk-averse linear estimation. Let $y = Ax + w$ with $w \sim \mathcal{N}(0, I)$ and $A \in \mathbf{R}^{q\times p}$. Here x is an unknown quantity that we wish to estimate, y is the measurement, and w is the measurement noise. We assume that $p \leq q$ and that A has full column rank.

A linear estimator $\widehat{x} = My$, with $M \in \mathbf{R}^{p\times q}$, is unbiased if $\mathbf{E}\widehat{x} = x$ where $\mathbf{E}$ means expected value, i.e., the estimator is unbiased if $MA = I$. The minimum-variance unbiased estimator is the unbiased estimator that minimizes the error variance

$$\mathbf{E}\|My - x\|^2 = \mathbf{Tr}MM^T = \sum_{i=1}^{p} \sigma_i^2(M),$$

where $\sigma_i(M)$ is the ith largest singular value of M. It is given by $M = A^+$, where $A^+ = (A^TA)^{-1}A^T$ is the pseudoinverse of A. In fact the minimum-variance estimator is optimal in a stronger sense: it not only minimizes $\sum_i \sigma_i^2(M)$ but each singular value $\sigma_i(M)$ separately:

$$MA = I \Longrightarrow \sigma_i(A^+) \leq \sigma_i(M), \quad i = 1, \ldots, p. \tag{2.6}$$

In some applications estimation errors larger than the mean value are more costly, or less desirable, than errors less than the mean value. To capture this idea of *risk aversion* we can consider the objective or cost function

$$2\gamma^2 \log \mathbf{E} \exp\left(\frac{1}{2\gamma^2}\|My - x\|^2\right), \tag{2.7}$$

where the parameter γ is called the *risk-sensitivity parameter*. This cost function was introduced by Whittle in the more sophisticated setting of stochastic optimal control; see [72, sect. 19]. Note that as $\gamma \to \infty$, the risk-sensitive cost (2.7) converges to the cost $\mathbf{E}\|My - x\|^2$, and is always larger (by convexity of exp). We can gain further insight from the first terms of the series expansion in $1/\gamma^2$:

$$\begin{aligned} 2\gamma^2 \log \mathbf{E} \exp\left(\frac{1}{2\gamma^2}\|\widehat{x} - x\|^2\right) &\simeq \mathbf{E}\|\widehat{x} - x\|^2 + \frac{1}{4\gamma^2}\left(\mathbf{E}\|\widehat{x} - x\|^4 - \left(\mathbf{E}\|\widehat{x} - x\|^2\right)^2\right) \\ &= \mathbf{E}z + \frac{1}{4\gamma^2}\mathrm{var}\ z, \end{aligned}$$

where $z = \|\widehat{x} - x\|^2$ is the squared error. Thus, for large γ, the risk-averse cost (2.7) augments the mean-square error with a term proportional to the variance of the squared error.

The unbiased, risk-averse optimal estimator can be found by solving

$$\begin{array}{ll} \text{minimize} & 2\gamma^2 \log \mathbf{E} \exp\left(\frac{1}{2\gamma^2}\|My - x\|^2\right) \\ \text{subject to} & MA = I, \end{array}$$

which can be expressed as a max-det problem. The objective function can be written as

$$\begin{aligned} &2\gamma^2 \log \mathbf{E} \exp\left(\frac{1}{2\gamma^2}\|My - x\|^2\right) \\ &= 2\gamma^2 \log \mathbf{E} \exp\left(\frac{1}{2\gamma^2} w^T M^T M w\right) \end{aligned}$$

$$= \begin{cases} 2\gamma^2 \log\det(I - (1/\gamma^2)M^TM)^{-1/2} & \text{if } M^TM \prec \gamma^2 I, \\ \infty & \text{otherwise,} \end{cases}$$

$$= \begin{cases} \gamma^2 \log\det \begin{bmatrix} I & \gamma^{-1}M^T \\ \gamma^{-1}M & I \end{bmatrix}^{-1} & \text{if } \begin{bmatrix} I & \gamma^{-1}M^T \\ \gamma^{-1}M & I \end{bmatrix} \succ 0, \\ \infty & \text{otherwise,} \end{cases}$$

so the unbiased risk-averse optimal estimator solves the max-det problem

$$\begin{array}{ll} \text{minimize} & \gamma^2 \log\det \begin{bmatrix} I & \gamma^{-1}M^T \\ \gamma^{-1}M & I \end{bmatrix}^{-1} \\ \text{subject to} & \begin{bmatrix} I & \gamma^{-1}M^T \\ \gamma^{-1}M & I \end{bmatrix} \succ 0 \\ & MA = I. \end{array} \tag{2.8}$$

This is in fact an analytic centering problem, and has a simple analytic solution: the least squares estimator $M = A^+$. To see this we express the objective in terms of the singular values of M:

$$\gamma^2 \log\det \begin{bmatrix} I & \gamma^{-1}M^T \\ \gamma^{-1}M & I \end{bmatrix}^{-1} = \begin{cases} \gamma^2 \displaystyle\sum_{i=1}^{p} \log(1 - \sigma_i^2(M)/\gamma^2)^{-1} & \text{if } \sigma_1(M) < \gamma, \\ \infty & \text{otherwise.} \end{cases}$$

It follows from property (2.6) that the solution is $M = A^+$ if $\|A^+\| < \gamma$, and that the problem is infeasible otherwise. (Whittle refers to the infeasible case, in which the risk-averse cost is always infinite, as "neurotic breakdown.")

In the simple case discussed above, the optimal risk-averse and the minimum-variance estimators coincide (so there is certainly no advantage in a max-det problem formulation). When additional convex constraints on the matrix M are added, e.g., a given sparsity pattern, or triangular or Toeplitz structure, the optimal risk-averse estimator can be found by including these constraints in the max-det problem (2.8) (and will not, in general, coincide with the constrained minimum-variance estimator).

2.4. Experiment design.

Optimal experiment design. As in the previous section, we consider the problem of estimating a vector x from a measurement $y = Ax + w$, where $w \sim \mathcal{N}(0, I)$ is measurement noise. The error covariance of the minimum-variance estimator is equal to $A^+(A^+)^T = (A^TA)^{-1}$. We suppose that the rows of the matrix $A = [a_1 \ \ldots \ a_q]^T$ can be chosen among M possible test vectors $v^{(i)} \in \mathbf{R}^p$, $i = 1, \ldots, M$:

$$a_i \in \{v^{(1)}, \ldots, v^{(M)}\}, \quad i = 1, \ldots, q.$$

The goal of experiment design is to choose the vectors a_i so that the error covariance $(A^TA)^{-1}$ is "small." We can interpret each component of y as the result of an experiment or measurement that can be chosen from a fixed menu of possible experiments; our job is to find a set of measurements that (together) are maximally informative.

We can write $A^TA = q\sum_{i=1}^{M} \lambda_i v^{(i)} {v^{(i)}}^T$, where λ_i is the fraction of rows a_k equal to the vector $v^{(i)}$. We ignore the fact that the numbers λ_i are integer multiples of $1/q$, and instead treat them as continuous variables, which is justified in practice when q

is large. (Alternatively, we can imagine that we are designing a random experiment: each experiment a_i has the form $v^{(k)}$ with probability λ_k.)

Many different criteria for measuring the size of the matrix $(A^TA)^{-1}$ have been proposed. For example, in E-optimal design, we minimize the norm of the error covariance, $\lambda_{\max}((A^TA)^{-1})$, which is equivalent to maximizing the smallest eigenvalue of A^TA. This is readily cast as the SDP

$$\begin{array}{ll} \text{maximize} & t \\ \text{subject to} & \displaystyle\sum_{i=1}^M \lambda_i v^{(i)} {v^{(i)}}^T \succeq tI \\ & \displaystyle\sum_{i=1}^M \lambda_i = 1 \\ & \lambda_i \geq 0, \quad i = 1, \ldots, M, \end{array}$$

in the variables $\lambda_1, \ldots, \lambda_M$, and t. Another criterion is A-optimality, in which we minimize $\mathbf{Tr}(A^TA)^{-1}$. This can be cast as an SDP:

$$\begin{array}{ll} \text{minimize} & \displaystyle\sum_{i=1}^p t_i \\ \text{subject to} & \left[\begin{array}{cc} \sum_{i=1}^M \lambda_i v^{(i)} {v^{(i)}}^T & e_i \\ e_i^T & t_i \end{array} \right] \succeq 0, \quad i = 1, \ldots, p, \\ & \lambda_i \geq 0, \quad i = 1, \ldots, M, \\ & \displaystyle\sum_{i=1}^M \lambda_i = 1, \end{array}$$

where e_i is the ith unit vector in $\mathbf{R}^p$, and the variables are λ_i, $i = 1, \ldots, M$, and t_i, $i = 1, \ldots, p$.

In D-optimal design, we minimize the determinant of the error covariance $(A^TA)^{-1}$, which leads to the max-det problem

$$\begin{array}{ll} \text{minimize} & \log\det\left(\displaystyle\sum_{i=1}^M \lambda_i v^{(i)} {v^{(i)}}^T\right)^{-1} \\ \text{subject to} & \lambda_i \geq 0, \ i = 1, \ldots, M \\ & \displaystyle\sum_{i=1}^M \lambda_i = 1. \end{array} \tag{2.9}$$

In section 3 we will derive an interesting geometrical interpretation of the D-optimal matrix A, and show that A^TA determines the minimum volume ellipsoid, centered at the origin, that contains $v^{(1)}, \ldots, v^{(M)}$.

Fedorov [33], Atkinson and Donev [7], Pukelsheim [55], and Cook and Fedorov [19] give surveys and additional references on optimal experiment design. Wilhelm [74], [73] discusses nondifferentiable optimization methods for experiment design. Jávorzky et al. [40] describe an application in frequency domain system identification, and compare the interior-point method discussed later in this paper with conventional algorithms. Ko, Lee, and Wayne [44], Lee [45], and Anstreicher et al. [5] discuss a nonconvex experiment design problem and a relaxation solved by an interior-point method.

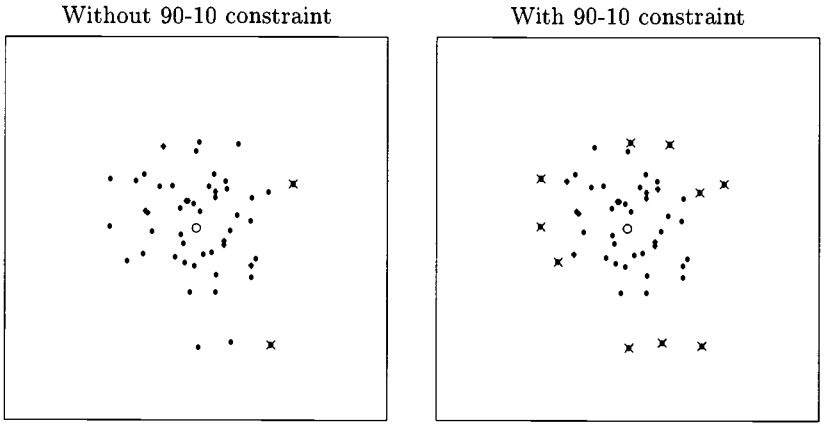


FIG. 2.1. *A D-optimal experiment design involving* 50 *test vectors in* $\mathbf{R}^2$*, with and without the* 90-10 *constraint. The circle is the origin; the dots are the test vectors that are not used in the experiment (i.e., have a weight* $\lambda_i = 0$*); the crosses are the test vectors that are used (i.e., have a weight* $\lambda_i > 0$*). Without the* 90-10 *constraint, the optimal design allocates all measurements to only two test vectors. With the constraint, the measurements are spread over* 10 *vectors, with no more than* 90% *of the measurements allocated to any group of five vectors. See also Figure* 2.2.

Extensions of *D*-optimal experiment design. The formulation of D-optimal design as a max-det problem has the advantage that one can easily incorporate additional useful convex constraints. For example, one can add linear inequalities $c_i^T \lambda \leq \alpha_i$, which can reflect bounds on the total cost of, or time required to carry out, the experiments.

We can also consider the case where each experiment yields several measurements, i.e., the vectors a_i and $v^{(k)}$ become matrices. The max-det problem formulation (2.9) remains the same, except that the terms $v^{(k)}v^{(k)T}$ can now have rank larger than one. This extension is useful in conjunction with additional linear inequalities representing limits on cost or time: we can model discounts or time savings associated with performing groups of measurements simultaneously. Suppose, for example, that the cost of simultaneously making measurements $v^{(1)}$ and $v^{(2)}$ is less than the sum of the costs of making them separately. We can take $v^{(3)}$ to be the matrix

$$v^{(3)} = \begin{bmatrix} v^{(1)} & v^{(2)} \end{bmatrix}$$

and assign costs c_1, c_2, and c_3 associated with making the first measurement alone, the second measurement alone, and the two simultaneously, respectively.

Let us describe in more detail another useful additional constraint that can be imposed: that no more than a certain fraction of the total number of experiments, say 90%, is concentrated in less than a given fraction, say 10%, of the possible measurements. Thus, we require

$$\sum_{i=1}^{\lfloor M/10 \rfloor} \lambda_{[i]} \leq 0.9, \tag{2.10}$$

where $\lambda_{[i]}$ denotes the ith largest component of λ. The effect on the experiment design will be to spread out the measurements over more points (at the cost of increasing the determinant of the error covariance). See Figures 2.1 and 2.2.

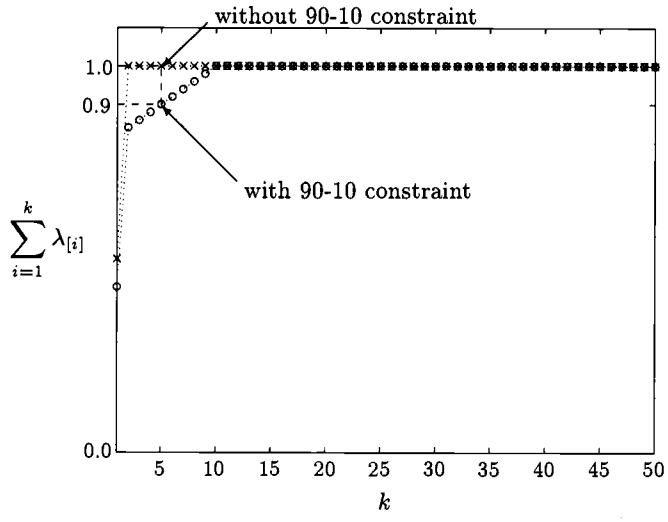


FIG. 2.2. *Experiment design of Figure* 2.1. *The curves show the sum of the largest* k *components of* λ *as a function of* k*, without the* 90-10 *constraint ("×"), and with the constraint ("∘"). The constraint specifies that the sum of the largest five components should be less than* 0.9*, i.e., the curve should avoid the area inside the dashed rectangle.*

The constraint (2.10) is convex; it is satisfied if and only if there exists $x \in \mathbf{R}^M$ and t such that

$$\begin{aligned} &\lfloor M/10 \rfloor\, t + \sum_{i=1}^{M} x_i \le 0.9, \\ &t + x_i \ge \lambda_i, \quad i = 1, \ldots, M, \\ &x \ge 0 \end{aligned} \tag{2.11}$$

(see [14, p. 318]). One can therefore compute the D-optimal design subject to the 90-10 constraint (2.10) by adding the linear inequalities (2.11) to the constraints in (2.9) and solving the resulting max-det problem in the variables λ, x, t.

2.5. Maximum likelihood estimation of structured covariance matrices. The next example is the maximum likelihood (ML) estimation of structured covariance matrices of a normal distribution. This problem has a long history; see, e.g., Anderson [2], [3].

Let $y^{(1)}, \ldots, y^{(M)}$ be M samples from a normal distribution $\mathcal{N}(0, \Sigma)$. The ML estimate for Σ is the positive definite matrix that maximizes the log-likelihood function $\log \prod_{i=1}^{M} p(y^{(i)})$, where

$$p(x) = ((2\pi)^p \det \Sigma)^{-1/2} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right).$$

In other words, Σ can be found by solving

$$\begin{array}{ll} \text{maximize} & \log\det \Sigma^{-1} - \frac{1}{M} \displaystyle\sum_{i=1}^{N} {y^{(i)}}^T \Sigma^{-1} y^{(i)} \\ \text{subject to} & \Sigma \succ 0. \end{array} \tag{2.12}$$

This can be expressed as a max-det problem in the *inverse* $R = \Sigma^{-1}$:

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}SR + \log\det R^{-1} \\ \text{subject to} & R \succ 0, \end{array} \tag{2.13}$$

where $S = \frac{1}{M}\sum_{i=1}^{n} y^{(i)} y^{(i)^T}$. Problem (2.13) has the straightforward analytical solution $R = S^{-1}$ (provided S is nonsingular).

It is often useful to impose additional structure on the covariance matrix Σ or its inverse R (Anderson [2], [3], Burg, Luenberger, and Wenger [15], Scharf [61, sect. 6.13], Dembo [25]). In some special cases (e.g., Σ is circulant) analytical solutions are known; in other cases where the constraints can be expressed as LMIs in R, the ML estimate can be obtained from a max-det problem. To give a simple illustration, bounds on the variances Σ_{ii} can be expressed as LMIs in R:

$$\Sigma_{ii} = e_i^T R^{-1} e_i \leq \alpha \Longleftrightarrow \left[\begin{array}{cc} R & e_i \\ e_i^T & \alpha \end{array}\right] \succeq 0.$$

The formulation as a max-det problem is also useful when the matrix S is singular (for example, because the number of samples is too small) and, as a consequence, the max-det problem (2.13) is unbounded below. In this case we can impose constraints (i.e., prior information) on Σ, for example, lower and upper bounds on the diagonal elements of R.

2.6. Gaussian channel capacity.

The Gaussian channel and the water-filling algorithm. The entropy of a normal distribution $\mathcal{N}(\mu, \Sigma)$ is, up to a constant, equal to $\frac{1}{2}\log\det\Sigma$ (see Cover and Thomas [21, Chap. 9]). It is therefore not surprising that max-det problems arise naturally in information theory and communications. One example is the computation of channel capacity.

Consider a simple Gaussian communication channel: $y = x + v$, where y, x, and v are random vectors in $\mathbf{R}^n$; $x \sim \mathcal{N}(0, X)$ is the input; y is the output, and $v \sim \mathcal{N}(0, R)$ is additive noise, independent of x. This model can represent n parallel channels, or one single channel at n different time instants or n different frequencies.

We assume the noise covariance R is known and given; the input covariance X is the variable to be determined, subject to constraints (such as power limits) that we will describe below. Our goal is to maximize the *mutual information* between input and output, given by

$$\frac{1}{2}\left(\log\det(X + R) - \log\det R\right) = \frac{1}{2}\log\det(I + R^{-1/2} X R^{-1/2})$$

(see [21]). The *channel capacity* is defined as the maximum mutual information over all input covariances X that satisfy the constraints. (Thus, the channel capacity depends on R and the constraints.)

The simplest and most common constraint is a limit on the average total power in the input, i.e.,

$$\mathbf{E}x^T x/n = \mathbf{Tr}X/n \leq P. \tag{2.14}$$

The information capacity subject to this average power constraint is the optimal value of

$$\begin{array}{ll} \text{maximize} & \frac{1}{2}\log\det(I + R^{-1/2} X R^{-1/2}) \\ \text{subject to} & \mathbf{Tr}X \leq nP \\ & X \succeq 0 \end{array} \tag{2.15}$$

(see [21, sect. 10]). This is a max-det problem in the variable $X = X^T$.

There is a straightforward solution to (2.15), known in information theory as the *water-filling* algorithm (see [21, sect. 10], [20]). Let $R = V\Lambda V^T$ be the eigenvalue decomposition of R. By introducing a new variable $\widetilde{X} = V^T X V$, we can rewrite the problem as

$$\begin{array}{ll} \text{maximize} & \frac{1}{2}\log\det(I + \Lambda^{-1/2}\widetilde{X}\Lambda^{-1/2}) \\ \text{subject to} & \mathbf{Tr}\widetilde{X} \leq nP \\ & \widetilde{X} \succeq 0. \end{array}$$

Since the off-diagonal elements of $\widetilde{X}$ do not appear in the constraints, but decrease the objective, the optimal $\widetilde{X}$ is diagonal. Using Lagrange multipliers one can show that the solution is $\widetilde{X}_{ii} = \max(\nu - \lambda_i, 0)$, $i = 1, \ldots, n$, where the Lagrange multiplier ν is to be determined from $\sum \widetilde{X}_{ii} = nP$. The term "water filling" refers to a visual description of this procedure (see [21, sect. 10], [20]).

Average power constraints on each channel. Problem (2.15) can be extended and modified in many ways. For example, we can replace the average *total* power constraint by an average power constraint on the individual channels, i.e., we can replace (2.14) by $\mathbf{E}x_k^2 = X_{kk} \leq P$, $k = 1, \ldots, n$. The capacity subject to this constraint can be determined by solving the max-det problem

$$\begin{array}{ll} \text{maximize} & \frac{1}{2}\log\det\left(I + R^{-1/2}XR^{-1/2}\right) \\ \text{subject to} & X \succeq 0 \\ & X_{kk} \leq P, \quad k = 1, \ldots, n. \end{array}$$

The water-filling algorithm does not apply here, but the capacity is readily computed by solving this max-det problem in X. Moreover, we can easily add other constraints, such as power limits on subsets of individual channels, or an upper bound on the correlation coefficient between two components of x:

$$\frac{|X_{ij}|}{\sqrt{X_{ii}X_{jj}}} \leq \rho_{\max} \Longleftrightarrow \begin{bmatrix} \sqrt{\rho_{\max}}X_{ii} & X_{ij} \\ X_{ij} & \sqrt{\rho_{\max}}X_{jj} \end{bmatrix} \succeq 0.$$

Gaussian channel capacity with feedback. Suppose that the n components of x, y, and v are consecutive values in a time series. The question whether knowledge of the past values v_k helps in increasing the capacity of the channel is of great interest in information theory [21, sect. 10.6]). In the Gaussian channel with feedback one uses, instead of x, the vector $\tilde{x} = Bv + x$ as input to the channel, where B is a strictly lower triangular matrix. The output of the channel is $y = \tilde{x} + v = x + (B + I)v$. We assume there is an average total power constraint: $\mathbf{E}\tilde{x}^T\tilde{x}/n \leq P$.

The mutual information between $\tilde{x}$ and y is

$$\frac{1}{2}\left(\log\det((B+I)R(B+I)^T + X) - \log\det R\right),$$

so we maximize the mutual information by solving

$$\begin{array}{ll} \text{maximize} & \frac{1}{2}\left(\log\det((B+I)R(B+I)^T + X) - \log\det R\right) \\ \text{subject to} & \mathbf{Tr}(BRB^T + X) \leq nP \\ & X \succeq 0 \\ & B \text{ strictly lower triangular} \end{array}$$

over the matrix variables B and X. To cast this problem as a max-det problem, we introduce a new variable $Y = (B+I)R(B+I)^T + X$ (i.e., the covariance of y), and obtain

$$\begin{array}{ll} \text{maximize} & \log\det Y \\ \text{subject to} & \mathbf{Tr}(Y - RB^T - BR - R) \le nP \\ & Y - (B+I)R(B+I)^T \succeq 0 \\ & B \text{ strictly lower triangular.} \end{array} \tag{2.16}$$

The second constraint can be expressed as an LMI in B and Y,

$$\left[\begin{array}{cc} Y & B+I \\ (B+I)^T & R^{-1} \end{array}\right] \succeq 0,$$

so (2.16) is a max-det problem in B and Y.

Capacity of channel with crosstalk. Suppose the n channels are independent, i.e., all covariances are diagonal, and that the noise covariance depends on X: $R_{ii} = r_i + a_i X_{ii}$, with $a_i > 0$. This has been used as a model of near-end crosstalk (see [6]). The capacity (with the total average power constraint) is the optimal value of

$$\begin{array}{ll} \text{maximize} & \dfrac{1}{2}\displaystyle\sum_{i=1}^{n} \log\left(1 + \frac{X_{ii}}{r_i + a_i X_{ii}}\right) \\ \text{subject to} & X_{ii} \ge 0, \quad i = 1, \ldots, n, \\ & \displaystyle\sum_{i=1}^{n} X_{ii} \le nP, \end{array}$$

which can be cast as a max-det problem

$$\begin{array}{ll} \text{maximize} & \frac{1}{2}\displaystyle\sum_{i=1}^{n} \log(1+t_i) \\ \text{subject to} & X_{ii} \ge 0,\ t_i \ge 0, \quad i = 1, \ldots, n, \\ & \left[\begin{array}{cc} 1 - a_i t_i & \sqrt{r}_i \\ \sqrt{r}_i & a_i X_{ii} + r_i \end{array}\right] \succeq 0, \quad i = 1, \ldots, n, \\ & \displaystyle\sum_{i=1}^{n} X_{ii} \le nP. \end{array}$$

The LMI is equivalent to $t_i \le X_{ii}/(r_i + a_i X_{ii})$. This problem can be solved using standard methods; the advantage of a max-det problem formulation is that we can add other (LMI) constraints on X, e.g., individual power limits. As another interesting possibility, we could impose constraints that distribute the power across the channels more uniformly, e.g., a 90-10 type constraint (see section 2.4).

3. The dual problem. We associate with (1.1) the *dual* problem

$$\begin{array}{ll} \text{maximize} & \log\det W - \mathbf{Tr} G_0 W - \mathbf{Tr} F_0 Z + l \\ \text{subject to} & \mathbf{Tr} G_i W + \mathbf{Tr} F_i Z = c_i, \quad i = 1, ..., m, \\ & W = W^T \succ 0, \quad Z = Z^T \succeq 0. \end{array} \tag{3.1}$$

The variables are $W \in \mathbf{R}^{l\times l}$ and $Z \in \mathbf{R}^{n\times n}$. Problem (3.1) is also a max-det problem, and can be converted into a problem of the form (1.1) by elimination of the equality constraints.

We say W and Z are dual feasible if they satisfy the constraints in (3.1), and strictly dual feasible if in addition $Z \succ 0$. We also refer to the max-det problem (1.1) as the primal problem and say x is primal feasible if $F(x) \succeq 0$ and $G(x) \succ 0$, and strictly primal feasible if $F(x) \succ 0$ and $G(x) \succ 0$.

Let p^* and d^* be the optimal values of problem (1.1) and (3.1), respectively (with the convention that $p^* = +\infty$ if the primal problem is infeasible, and $d^* = -\infty$ if the dual problem is infeasible).

The optimization problem (3.1) is the Lagrange dual of problem (1.1), rewritten as

$$\begin{array}{ll} \text{minimize} & c^T x + \log\det X^{-1} \\ \text{subject to} & X = G(x) \\ & F(x) \succeq 0, \quad X \succ 0. \end{array}$$

(We introduce a new variable $X = X^T \in \mathbf{R}^{l\times l}$ and add an equality constraint.) To derive the dual problem, we associate a Lagrange multiplier $Z = Z^T \succeq 0$ with the LMI $F(x) \succeq 0$, and a multiplier $W = W^T$ with the equality constraint $X = G(x)$. The optimal value can then be expressed as

$$p^* = \inf_{x,X} \sup_{Z\succeq 0, W} \left(c^T x + \log\det X^{-1} - \mathbf{Tr} Z F(x) + \mathbf{Tr} W(X - G(x)) \right).$$

Changing the order of the supremum and the infimum, and solving the inner unconstrained minimization over x and X analytically, yields a lower bound on p^*:

$$\begin{aligned} p^* &\geq \sup_{Z\succeq 0, W} \inf_{x,X} \left(c^T x + \log\det X^{-1} - \mathbf{Tr} Z F(x) + \mathbf{Tr} W(X - G(x)) \right) \\ &= \sup_{Z\succeq 0,\, W\succ 0,\, c_i = \mathbf{Tr} Z F_i + \mathbf{Tr} W G_i} (\log\det W - \mathbf{Tr} Z F_0 - \mathbf{Tr} W G_0 + l) \\ &= d^*. \end{aligned}$$

The inequality $p^* \geq d^*$ holds with equality if a constraint qualification holds, as stated in the following theorem.

THEOREM 3.1. *$p^* \geq d^*$. If (1.1) is strictly feasible, the dual optimum is achieved; if (3.1) is strictly feasible, the primal optimum is achieved. In both cases, $p^* = d^*$.*

The theorem follows from standard results in convex optimization (Luenberger [48, Chap. 8], Rockafellar [57, sect. 29–30], Hiriart-Urruty and Lemaréchal [39, Chap. XII]), so we will not prove it here. See also Lewis [46] for a more general discussion of convex analysis of functions of symmetric matrices.

The difference between the primal and dual objective, i.e., the expression

$$\begin{aligned} & c^T x + \log\det G(x)^{-1} + \log\det W^{-1} + \mathbf{Tr} G_0 W + \mathbf{Tr} F_0 Z - l \\ &= \sum_{i=1}^m x_i \mathbf{Tr} G_i W + \mathbf{Tr} G_0 W + \sum_{i=1}^m x_i \mathbf{Tr} F_i Z + \mathbf{Tr} F_0 Z - \log\det G(x) W - l \\ &= \mathbf{Tr} G(x) W - \log\det G(x) W - l + \mathbf{Tr} F(x) Z, \end{aligned} \tag{3.2}$$

is called the *duality gap* associated with x, W, and Z. Theorem 3.1 states that the duality gap is always nonnegative, and zero only if x, W, and Z are optimal.

Note that zero duality gap (3.2) implies $G(x)W = I$ and $F(x)Z = 0$. This gives the optimality condition for the max-det problem (1.1): a primal feasible x is optimal if there exists a $Z \succeq 0$, such that $F(x)Z = 0$ and

$$\mathbf{Tr} G_i G(x)^{-1} + \mathbf{Tr} F_i Z = c_i, \quad i = 1, \ldots, m.$$

This optimality condition is always sufficient; it is also necessary if the primal problem is strictly feasible.

In the remainder of the paper we will assume that the max-det problem is strictly primal and dual feasible. By Theorem 3.1, this assumption implies that the primal problem is bounded below and the dual problem is bounded above, with equality at the optimum, and that the primal and dual optimal sets are nonempty.

Example. Semidefinite programming dual. As an illustration, we derive from (3.1) the dual problem for the SDP (1.2). Substituting $G_0 = 1$, $G_i = 0$, $l = 1$, in (3.1) yields

$$\begin{array}{ll} \text{maximize} & \log W - W - \mathbf{Tr} F_0 Z + 1 \\ \text{subject to} & \mathbf{Tr} F_i Z = c_i, \quad i = 1, \ldots, m, \\ & W \succ 0, \quad Z \succeq 0. \end{array}$$

The optimal value of W is one, so the dual problem reduces to

$$\begin{array}{ll} \text{maximize} & -\mathbf{Tr} F_0 Z \\ \text{subject to} & \mathbf{Tr} F_i Z = c_i, \quad i = 1, \ldots, m, \\ & Z \succeq 0, \end{array}$$

which is the dual SDP (in the notation used in [69]).

Example. D-optimal experiment design. As a second example we derive the dual of the experiment design problem (2.9). After a few simplifications we obtain

$$\begin{array}{ll} \text{maximize} & \log\det W + p - z \\ \text{subject to} & W = W^T \succ 0 \\ & {v^{(i)}}^T W v^{(i)} \leq z, \ i = 1, \ldots, M, \end{array} \tag{3.3}$$

where the variables are the matrix W and the scalar variable z. Problem (3.3) can be further simplified. The constraints are homogeneous in W and z, so for each dual feasible W, z we have a ray of dual feasible solutions tW, tz, $t > 0$. It turns out that we can analytically optimize over t: replacing W by tW and z by tz changes the objective to $\log\det W + p\log t + p - tz$, which is maximized for $t = p/z$. After this simplification, and with a new variable $\widetilde{W} = (p/z)W$, problem (3.3) becomes

$$\begin{array}{ll} \text{maximize} & \log\det \widetilde{W} \\ \text{subject to} & \widetilde{W} \succ 0 \\ & {v^{(i)}}^T \widetilde{W} v^{(i)} \leq p, \ i = 1, \ldots, M. \end{array} \tag{3.4}$$

Problem (3.4) has an interesting geometrical meaning: the constraints state that $\widetilde{W}$ determines an ellipsoid $\{x \mid x^T \widetilde{W} x \leq p\}$, centered at the origin, that contains the points $v^{(i)}$, $i = 1, \ldots, M$; the objective is to maximize $\det \widetilde{W}$, i.e., to minimize the volume of the ellipsoid.

There is an interesting connection between the optimal primal variables λ_i and the points $v^{(i)}$ that lie on the boundary of the optimal ellipsoid $\mathcal{E}$. First, note that

the duality gap associated with a primal feasible λ and a dual feasible $\widetilde{W}$ is equal to

$$\log\det\left(\sum_{i=1}^{M}\lambda_i v^{(i)}v^{(i)^T}\right)^{-1} - \log\det\widetilde{W},$$

and is zero (hence λ is optimal) if and only if $\widetilde{W} = \left(\sum_{i=1}^{M}\lambda_i v^{(i)}v^{(i)^T}\right)^{-1}$. Hence λ is optimal if

$$\mathcal{E} = \left\{ x \in \mathbf{R}^p \;\middle|\; x^T\left(\sum_{i=1}^{M}\lambda_i v^{(i)}v^{(i)^T}\right)^{-1} x \le p \right\}$$

is the minimum volume ellipsoid, centered at the origin, that contains the points $v^{(j)}$, $j = 1, \ldots, M$. We also have (in fact, for any feasible λ)

$$\begin{aligned}
&\sum_{j=1}^{M}\lambda_j\left(p - v^{(j)^T}\left(\sum_{i=1}^{M}\lambda_i v^{(i)}v^{(i)^T}\right)^{-1} v^{(j)}\right)\\
&= p - \mathbf{Tr}\left(\sum_{j=1}^{M}\lambda_j v^{(j)}v^{(j)^T}\right)\left(\sum_{i=1}^{M}\lambda_i v^{(i)}v^{(i)^T}\right)^{-1}\\
&= 0.
\end{aligned}$$

If λ is optimal, then each term in the sum on the left-hand side is positive (since $\mathcal{E}$ contains all vectors $v^{(j)}$), and therefore the sum can only be zero if each term is zero:

$$\lambda_j > 0 \Longrightarrow v^{(j)^T}\left(\sum_{i=1}^{M}\lambda_i v^{(i)}v^{(i)^T}\right)^{-1} v^{(j)} = p.$$

Geometrically, λ_j is nonzero only if $v^{(j)}$ lies on the boundary of the minimum volume ellipsoid. This makes more precise the intuitive idea that an optimal experiment only uses "extreme" test vectors. Figure 3.1 shows the optimal ellipsoid for the experiment design example of Figure 2.1.

The duality between D-optimal experiment designs and minimum volume ellipsoids also extends to nonfinite compacts sets (Titterington [67], Pronzato and Walter [54]). The D-optimal experiment design problem on a compact set $C \subset \mathbf{R}^p$ is

$$\text{maximize} \;\; \log\det \mathbf{E}vv^T \tag{3.5}$$

over all probability measures on C. This is a convex but semi-infinite optimization problem, with dual [67]

$$\begin{array}{ll}
\text{maximize} & \log\det\widetilde{W}\\
\text{subject to} & \widetilde{W} \succ 0\\
& v^T\widetilde{W}v \le p, \;\; v \in C.
\end{array} \tag{3.6}$$

Again, we see that the dual is the problem of computing the minimum volume ellipsoid, centered at the origin, and covering the set C.

General methods for solving the semi-infinite optimization problems (3.5) and (3.6) fall outside the scope of this paper. In particular cases, however, these problems can

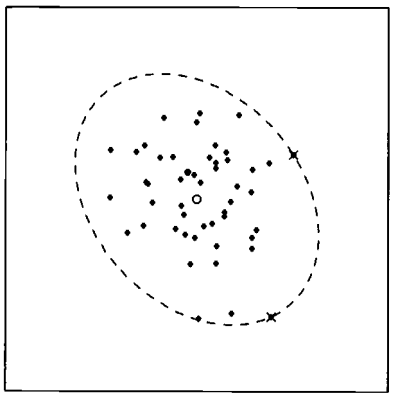


FIG. 3.1. *In the dual of the D-optimal experiment design problem we compute the minimum volume ellipsoid, centered at the origin, that contains the test vectors. The test vectors with a nonzero weight lie on the boundary of the optimal ellipsoid. Same data and notation as in Figure* 2.1.

be solved as max-det problems. One interesting example arises when C is the union of a finite number of ellipsoids. In this case, the dual (3.6) can be cast as a max-det problem (see [70]) and hence efficiently solved; by duality, we can recover from the dual solution the probability distribution that solves (3.5).

4. The central path. In this section we describe the *central path* of the max-det problem (1.1), and give some of its properties. The central path plays a key role in interior point methods for the max-det problem.

The primal central path. For strictly feasible x and $t \geq 1$, we define

$$\varphi_p(t,x) \stackrel{\Delta}{=} t\left(c^T x + \log\det G(x)^{-1}\right) + \log\det F(x)^{-1}. \tag{4.1}$$

This function is the sum of two convex functions: the first term is a positive multiple of the objective function in (1.1); the second term, $\log\det F(x)^{-1}$, is a barrier function for the set $\{x \mid F(x) \succ 0\}$. For future use, we note that the gradient and Hessian of $\varphi_p(x,t)$ are given by the expressions

$$\left(\nabla\varphi_p(t,x)\right)_i = t\left(c_i - \mathbf{Tr}G(x)^{-1}G_i\right) - \mathbf{Tr}F(x)^{-1}F_i, \tag{4.2}$$

$$\left(\nabla^2\varphi_p(t,x)\right)_{ij} = t\mathbf{Tr}G(x)^{-1}G_iG(x)^{-1}G_j + \mathbf{Tr}F(x)^{-1}F_iF(x)^{-1}F_j, \tag{4.3}$$

for $i,j = 1,\ldots,m$.

It can be shown that $\varphi_p(t,x)$ is a *strictly* convex function of x if the m matrices $\mathbf{diag}(G_i,F_i)$, $i = 1,\ldots,m$, are linearly independent, and that it is bounded below (since we assume the problem is strictly dual feasible). We define $x^\star(t)$ as the unique minimizer of $\varphi_p(t,x)$:

$$x^\star(t) = \operatorname{argmin}\left\{\varphi_p(t,x) \mid G(x) \succ 0, \;\; F(x) \succ 0\right\}.$$

The curve $x^\star(t)$, parameterized by $t \geq 1$, is called the *central path*.

The dual central path. Points $x^\star(t)$ on the central path are characterized by the optimality conditions $\nabla\varphi_p(t,x^\star(t)) = 0$, i.e., using the expression (4.2),

$$\mathbf{Tr}G(x^\star(t))^{-1}G_i + \frac{1}{t}\mathbf{Tr}F(x^\star(t))^{-1}F_i = c_i, \;\; i = 1,\ldots,m.$$

From this we see that the matrices

$$W^\star(t) = G(x^\star(t))^{-1}, \quad Z^\star(t) = \frac{1}{t}F(x^\star(t))^{-1} \tag{4.4}$$

are strictly dual feasible. The duality gap associated with $x^\star(t)$, $W^\star(t)$, and $Z^\star(t)$ is, from expression (3.2),

$$\mathbf{Tr}F(x^\star(t))Z^\star(t) + \mathbf{Tr}G(x^\star(t))W^\star(t) - \log\det G(x^\star(t))W^\star(t) - l = \frac{n}{t},$$

which shows that $x^\star(t)$ converges to the solution of the max-det problem as $t \to \infty$.

It can be shown that the pair $(W^\star(t), Z^\star(t))$ actually lies on the *dual* central path, defined as

$$(W^\star(t), Z^\star(t)) = \operatorname{argmin}\left\{ \varphi_d(t, W, Z) \;\middle|\; \begin{array}{l} W = W^T \succeq 0, \;\; Z = Z^T \succ 0, \\ \mathbf{Tr}G_i W + \mathbf{Tr}F_i Z = c_i, \;\; i = 1, \ldots, m \end{array} \right\},$$

where

$$\varphi_d(t, W, Z) \triangleq t\left(\log\det W^{-1} + \mathbf{Tr}G_0 W + \mathbf{Tr}F_0 Z - l\right) + \log\det Z^{-1}.$$

The close connections between primal and dual central path are summarized in the following theorem.

THEOREM 4.1. *If x is strictly primal feasible, and W, Z are strictly dual feasible, then*

$$\varphi_p(t, x) + \varphi_d(t, W, Z) \geq n(1 + \log t) \tag{4.5}$$

with equality if and only if $x = x^\star(t)$, $W = W^\star(t)$, $Z = Z^\star(t)$.

Proof. If $A = A^T \in \mathbf{R}^{p\times p}$ and $A \succ 0$, then $-\log\det A \geq -\mathbf{Tr}A + p$ (by convexity of $-\log\det A$ on the cone of positive semidefinite matrices). Applying this inequality, we find

$$\begin{aligned} \varphi_p(t, x) + \varphi_d(t, W, Z) &= t\left(\mathbf{Tr}G(x)W + \mathbf{Tr}F(x)Z - \log\det G(x)W - l\right) - \log\det F(x)Z \\ &= t\left(-\log\det W^{1/2}G(x)W^{1/2} + \mathbf{Tr}W^{1/2}G(x)W^{1/2}\right) \\ &\quad - \log\det tZ^{1/2}F(x)Z^{1/2} + \mathbf{Tr}tZ^{1/2}F(x)Z^{1/2} + n\log t - tl \\ &\geq tl + n + n\log t - tl = n(1 + \log t). \end{aligned}$$

The equality for $x = x^\star(t)$, $W = W^\star(t)$, $Z = Z^\star(t)$ can be verified by substitution.

Tangent to the central path. We conclude this section by describing how the tangent direction to the central path can be computed. Let $\phi_1(x) = -\log\det G(x)$ and $\phi_2(x) = -\log\det F(x)$. A point $x^\star(t)$ on the central path is characterized by

$$t\left(c + \nabla\phi_1(x^\star(t))\right) + \nabla\phi_2(x^\star(t)) = 0.$$

The tangent direction $\frac{\partial x^\star(t)}{\partial t}$ can be found by differentiating with respect to t:

$$c + \nabla\phi_1(x^\star(t)) + \left(t\nabla^2\phi_1(x^\star(t)) + \nabla^2\phi_2(x^\star(t))\right)\frac{\partial x^\star(t)}{\partial t} = 0,$$

so that

$$\frac{\partial x^\star(t)}{\partial t} = -\left(t\nabla^2\phi_1(x^\star(t)) + \nabla^2\phi_2(x^\star(t))\right)^{-1}\left(c + \nabla\phi_1(x^\star(t))\right). \tag{4.6}$$

By differentiating (4.4), we obtain the tangent to the dual central path,

$$\frac{\partial W^\star(t)}{\partial t} = -G(x^\star(t))^{-1}\left(\sum_{i=1}^m \frac{\partial x_i^\star(t)}{\partial t} G_i\right) G(x^\star(t))^{-1}, \tag{4.7}$$

$$\frac{\partial Z^\star(t)}{\partial t} = -\frac{1}{t^2}F(x^\star(t))^{-1} - \frac{1}{t}F(x^\star(t))^{-1}\left(\sum_{i=1}^m \frac{\partial x_i^\star(t)}{\partial t} F_i\right) F(x^\star(t))^{-1}. \tag{4.8}$$

5. Newton's method. In this section we consider the problem of minimizing $\varphi_p(t,x)$ for fixed t, i.e., computing $x^\star(t)$, given a strictly feasible initial point:

$$\begin{array}{ll} \text{minimize} & \varphi_p(t,x) \\ \text{subject to} & G(x) \succ 0 \\ & F(x) \succ 0. \end{array} \tag{5.1}$$

This includes, as a special case, the analytic centering problem ($t = 1$ and $F(x) = 1$). Our main motivation for studying (5.1) will become clear in the next section, when we discuss an interior-point method based on minimizing $\varphi_p(t,x)$ for a sequence of values t.

Newton's method with line search can be used to solve problem (5.1) efficiently.

Newton method for minimizing $\varphi_p(t,x)$
given strictly feasible x, tolerance δ $(0 < \delta \le 0.5)$
repeat

1. Compute the Newton direction $\delta x^N = -\left(\nabla^2\varphi_p(t,x)\right)^{-1}\nabla\varphi_p(t,x)$
2. Compute $\lambda = ((\delta x^N)^T\nabla^2\varphi_p(t,x)\delta x^N)^{1/2}$
3. **if** $(\lambda > 0.5)$, compute $\widehat{h} = \operatorname{argmin}\varphi_p(t, x + h\delta x^N)$
   else $\widehat{h} = 1$
4. Update: $x := x + \widehat{h}\delta x^N$

until $\lambda \le \delta$

The quantity

$$\lambda = ((\delta x^N)^T\nabla^2\varphi_p(t,x)\delta x^N)^{1/2} \tag{5.2}$$

is called the *Newton decrement* at x. The cost of step 3 (the *line search*) is very small, usually negligible compared with the cost of computing the Newton direction; see section 8 for details.

It is well known that the asymptotic convergence of Newton's method is quadratic. Nesterov and Nemirovsky in [51, sect. 2.2] give a complete analysis of the *global* speed of convergence. The main result of their convergence analysis applied to problem (5.1) is the following theorem.

THEOREM 5.1. *The algorithm terminates in fewer than*

$$11(\varphi_p(t,x^{(0)}) - \varphi_p(t,x^\star(t))) + \log_2\log_2(1/\delta) \tag{5.3}$$

iterations, and when it terminates, $\varphi_p(t,x) - \varphi_p(t,x^\star(t)) \le \delta$.

A self-contained proof is given in [70].

Note that the right-hand side of (5.3) does not depend on the problem size (i.e., m, n, or l) at all, and only depends on the problem data through the difference between the value of the function $\varphi_p(t,\cdot)$ at the initial point $x^{(0)}$ and at the central point $x^\star(t)$.

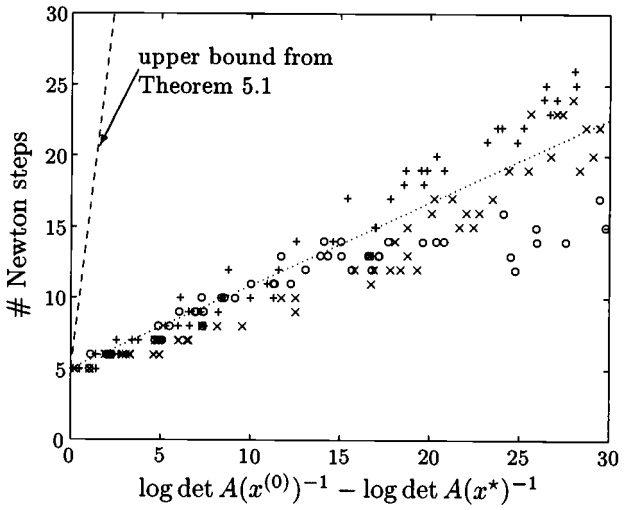


FIG. 5.1. *Number of Newton iterations to minimize* $\log\det A(x)^{-1}$ *versus* $\log\det A(x^{(0)})^{-1} - \log\det A(x^\star)^{-1}$ *(with* $\delta = 2.33\cdot 10^{-10}$*, i.e.,* $\log_2\log_2(1/\delta) = 5$*). Random matrix completion problems of three sizes ("+":* $m = 20$*;* $l = 20$*, "×":* $m = 100$*,* $l = 20$*, "∘":* $m = 20$*,* $l = 100$*). The dotted line is a least squares fit of the data and is given by* $5 + 0.59(\log\det A(x^{(0)})^{-1} - \log\det A(x^\star)^{-1})$*. The dashed line is the upper bound of Theorem* 5.1 *(*$5 + 11(\log\det A(x^{(0)})^{-1} - \log\det A(x^\star)^{-1})$*).*

The term $\log_2\log_2(1/\delta)$, which is characteristic of quadratic convergence, grows *extremely slowly* with required accuracy δ. For all practical purposes it can be considered a constant, say five (which guarantees an accuracy of $\delta = 2.33\cdot 10^{-10}$). Not quite precisely then, the theorem says we can compute $x^\star(t)$ in at most $11(\varphi_p(t, x^{(0)}) - \varphi_p(t, x^\star(t))) + 5$ Newton steps. The precise statement is that within this number of iterations we can compute an extremely good approximation of $x^\star(t)$. In what follows, we will speak of "computing the central point $x^\star(t)$" when we really mean computing an extremely good approximation. We can justify this on several grounds. It is possible to adapt our exposition to account for the extremely small approximation error incurred by terminating the Newton process after $11(\varphi_p(t, x^{(0)}) - \varphi_p(t, x^\star(t))) + 5$ steps. Indeed, the errors involved are certainly on the same scale as computer arithmetic (roundoff) errors, so if a complexity analysis is to be carried out with such precision, it should also account for roundoff error.

Theorem 5.1 holds for an "implementable" version of the algorithm as well, in which an appropriate *approximate* line search is used instead of the exact line search.

Numerical experiment. The bound provided by Theorem 5.1 on the number of Newton steps required to compute $x^\star(t)$, starting from $x^{(0)}$, will play an important role in our path-following method. It is therefore useful to examine how the bound compares to the actual number of Newton steps required in practice to compute $x^\star(t)$.

Figure 5.1 shows the results of a numerical experiment that compares the actual convergence of Newton's method with the bound (5.3). The test problem is a matrix completion problem

$$\begin{array}{ll} \text{minimize} & \log\det A(x)^{-1} \\ \text{subject to} & A(x) = A_f + \displaystyle\sum_{k=1}^{m} x_k \left(E_{i_k j_k} + E_{j_k i_k}\right) \succ 0, \end{array}$$

which is a particular case of (5.1) with $c = 0$, $G(x) = A(x)$, $F(x) = 1$, and $\varphi_p(t, x) = \log\det A(x)^{-1}$. We considered problems of three different sizes: $m = 20$, $l = 20$ (indicated by "+"); $m = 100$, $l = 20$ (indicated by "×"); $m = 20$, $l = 100$ (indicated by "∘"). Each point on the figure corresponds to a different problem instance, generated as follows.

- The matrices A_f were constructed as $A_f = UU^T$ with the elements of U drawn from a normal distribution $\mathcal{N}(0, 1)$. This guarantees that $x = 0$ is strictly feasible. The m index pairs (i_k, j_k), $i_k \neq j_k$, were chosen randomly with a uniform distribution over the off-diagonal index pairs. For each of the three problem sizes, 50 instances were generated.
- For each problem instance, we first computed $x^\star$ using $x = 0$ as starting point. We then selected a value γ (uniformly in the interval $(0, 30)$), generated a random $\widehat{x} \in \mathbf{R}^m$ (with distribution $\mathcal{N}(0, I)$), and then computed $x^{(0)} = x^\star + t(\widehat{x} - x^\star)$ such that

$$\log\det(A(x^{(0)}))^{-1} - \log\det(A(x^*))^{-1} = \gamma.$$

  This point $x^{(0)}$ was used as starting point for the Newton algorithm.

Our experience with other problems shows that the results for this family of random problems are quite typical.

From the results we can draw two important conclusions.

- The quantity $\log\det(A(x^{(0)}))^{-1} - \log\det(A(x^\star))^{-1}$ not only provides an upper bound on the number of Newton iterations via Theorem 5.1, but it is also a very good predictor of the number of iterations in practice. The dimensions m and l, on the other hand, have much less influence (except, of course, through $\log\det(A(x^{(0)}))^{-1} - \log\det(A(x^\star))^{-1}$).
- The average number of Newton iterations seems to grow as

$$\alpha + \beta\left(\log\det(A(x^{(0)}))^{-1} - \log\det(A(x^\star))^{-1}\right),$$

  with $\alpha \simeq 5$, $\beta \simeq 0.6$. This is significantly smaller than the upper bound of Theorem 5.1 ($\alpha = 5$, $\beta = 11$).

In summary, we conclude that the difference $\varphi_p(t, x^{(0)}) - \varphi_p(t, x^\star(t))$ is a good measure, in theory and in practice, of the effort required to compute $x^\star(t)$ using Newton's method, starting at $x^{(0)}$.

A computable upper bound on the number of Newton steps. Note that $\varphi_p(t, x^\star(t))$ is not known explicitly as a function of t. To evaluate the bound (5.3) one has to compute $x^\star(t)$, i.e., carry out the Newton algorithm (which, at the very least, would seem to defeat the purpose of trying to estimate or bound the number of Newton steps required to compute $x^\star(t)$). Therefore, the bound of Theorem 5.1 is not (directly) useful in practice. From Theorem 4.1, however, it follows that every dual feasible point W, Z provides a lower bound for $\varphi_p(t, x^\star(t))$:

$$\varphi_p(t, x^\star(t)) \geq -\varphi_d(t, W, Z) + n(1 + \log t)$$

and that the bound is exact if $W = W^\star(t)$ and $Z = Z^\star(t)$.

We can therefore replace the bound (5.3) by a weaker but more easily computed bound, provided we have a dual feasible pair W, Z:

$$\begin{aligned} &11(\varphi_p(t, x^{(0)}) - \varphi_p(t, x^\star(t))) + \log_2\log_2(1/\delta) \\ &\leq 11\psi_{\mathrm{ub}}(t, x^{(0)}, W, Z) + \log_2\log_2(1/\delta), \end{aligned} \tag{5.4}$$

where

$$\psi_{\rm ub}(t,x,W,Z) = \varphi_p(t,x) + \varphi_d(t,W,Z) - n(1+\log t). \tag{5.5}$$

This is the bound we will use in practice (and in our complexity analysis): it gives a readily computed bound on the number of Newton steps required to compute $x^\star(t)$, starting from $x^{(0)}$, given any dual feasible W, Z.

6. Path-following algorithms. Path-following methods for convex optimization have a long history. In their 1968 book [34], Fiacco and McCormick worked out many general properties, e.g., convergence to an optimal point, connections with duality, etc. No attempt was made to give a worst-case convergence analysis, until Renegar [56] proved polynomial convergence of a path-following algorithm for linear programming. Nesterov and Nemirovsky [51, sect. 3] studied the convergence for nonlinear convex problems and provided proofs of polynomial worst-case complexity. See [51, pp. 379–386] and den Hertog [28] for an historical overview.

We will present two variants of a path-following method for the max-det problem. The short-step version of section 6.2 is basically the path-following method of [34], [51], with a simplified, self-contained complexity analysis (see also Anstreicher and Fampa [4] for a very related analysis of an interior-point method for semidefinite programming). In the long-step version of section 6.3 we combine the method with predictor steps to accelerate convergence. This, too, is a well-known technique, originally proposed by Fiacco and McCormick; our addition is a new step selection rule.

6.1. General idea. One iteration proceeds as follows. The algorithm starts at a point $x^\star(t)$ on the central path. As we have seen above, the duality gap associated with $x^\star(t)$ is n/t. We then select a new value $t^+ > t$, and choose a strictly feasible starting point $\widehat{x}$ (which may or may not be equal to $x^\star(t)$). The point $\widehat{x}$ serves as an approximation of $x^\star(t^+)$ and is called the *predictor* of $x^\star(t^+)$. Starting at the predictor $\widehat{x}$, the algorithm computes $x^\star(t^+)$ using Newton's method. This reduces the duality gap by a factor t^+/t. The step from $x^\star(t)$ to $x^\star(t^+)$ is called an *outer iteration.*

The choice of t^+ and $\widehat{x}$ involves a trade-off. A large value of t^+/t means fast duality gap reduction, and hence fewer outer iterations. On the other hand, it makes it more difficult to find a good predictor $\widehat{x}$, and hence more Newton iterations may be needed to compute $x^\star(t^+)$.

In the method discussed below, we impose a bound on the maximum number of Newton iterations per outer iteration, by requiring that the predictor $\widehat{x}$ and the new value of t^+ satisfy

$$\varphi_p(t^+,\widehat{x}) - \varphi_p(t^+,x^\star(t^+)) \leq \gamma. \tag{6.1}$$

This implies that no more than $5 + 11\gamma$ Newton iterations are required to compute $x^\star(t^+)$ starting at $\widehat{x}$. Of course, the exact value of the left-hand side is not known, unless we carry out the Newton minimization, but as we have seen above, we can replace the condition by

$$\psi_{\rm ub}(t^+,\widehat{x},\widehat{W},\widehat{Z}) = \gamma, \tag{6.2}$$

where $\widehat{W}$ and $\widehat{Z}$ are conveniently chosen dual feasible points.

The parameters in the algorithm are $\gamma > 0$ and the desired accuracy ϵ.

Path-following algorithm

given $\gamma > 0$, $t \geq 1$, $x := x^\star(t)$
repeat
1. Select t^+, $\widehat{x}$, $\widehat{W}$, $\widehat{Z}$ such that $t^+ > t$ and $\psi_{\rm ub}(t^+, \widehat{x}, \widehat{W}, \widehat{Z}) = \gamma$
2. Compute $x^\star(t^+)$ starting at $\widehat{x}$, using the Newton algorithm of section 5
3. $t := t^+$, $x := x^\star(t^+)$

until $n/t \leq \epsilon$

Step 1 in this outline is not completely specified. In the next sections we will discuss in detail different choices. We will show that one can always find $\widehat{x}$, $\widehat{W}$, $\widehat{Z}$, and t^+ that satisfy

$$\frac{t^+}{t} \geq 1 + \sqrt{\frac{2\gamma}{n}}. \tag{6.3}$$

This fact allows us to estimate the total complexity of the method, i.e., to derive a bound on the total number of Newton iterations required to reduce the duality gap to ϵ. The algorithm starts on the central path, at $x^\star(t^{(0)})$, with initial duality gap $\epsilon^{(0)} = n/t^{(0)}$. Each iteration reduces the duality gap by t^+/t. Therefore, the total number of outer iterations required to reduce the initial gap of $\epsilon^{(0)}$ to a final value below ϵ is at most

$$\left\lceil \frac{\log(\epsilon^{(0)}/\epsilon)}{\log(1+\sqrt{2\gamma/n})} \right\rceil \leq \left\lceil \sqrt{n} \frac{\log(\epsilon^{(0)}/\epsilon)}{\log(1+\sqrt{2\gamma})} \right\rceil.$$

(The inequality follows from the concavity of $\log(1+x)$.) The total number of Newton steps can therefore be bounded as

$$\begin{aligned} \text{Total \#Newton iterations} &\leq \lceil 5 + 11\gamma \rceil \left\lceil \sqrt{n} \frac{\log(\epsilon^{(0)}/\epsilon)}{\log(1+\sqrt{2\gamma})} \right\rceil \\ &= O\left(\sqrt{n}\log(\epsilon^{(0)}/\epsilon)\right). \end{aligned} \tag{6.4}$$

This upper bound increases slowly with the problem dimensions: it grows as $\sqrt{n}$, and is independent of l and m. We will see later that the performance in practice is even better.

Note that we assume that the minimization in step 2 of the algorithm is exact. The justification of this assumption lies in the very fast local convergence of Newton's method: we have seen in section 5 that it takes only a few iterations to improve a solution with Newton decrement $\lambda \leq 0.5$ to one with a very high accuracy.

Nevertheless, in a practical implementation (as well as in a rigorous theoretical analysis), one has to take into account the fact that $x^\star(t)$ can only be computed approximately. For example, the stopping criterion $n/t \leq \epsilon$ is based on the duality gap associated with exactly central points $x^\star(t)$, $W^\star(t)$, and $Z^\star(t)$, and is therefore not quite accurate if $x^\star(t)$ is only known approximately. We give a suitably modified criterion in [70], where we show that dual feasible points are easily computed during the centering step (step 2) once the Newton decrement is less than one. Using the associated duality gap yields a completely rigorous stopping criterion. We will briefly point out some other modifications, as we develop different variants of the algorithm in the next sections; full details are described in [70]. With these modifications, the

algorithm works well even when $x^\star(t)$ is computed approximately. (We often use a value $\delta = 10^{-3}$ in the Newton algorithm.)

It is also possible to extend the simple worst-case complexity analysis to take into account incomplete centering, but we will not attempt such an analysis here. For the fixed-reduction algorithm (described immediately below), such a complete analysis can be found in Nesterov and Nemirovsky [51, sect. 3.2].

6.2. Fixed-reduction algorithm. The simplest variant uses $\widehat{x} = x^\star(t)$, $\widehat{W} = W^\star(t)$, and $\widehat{Z} = Z^\star(t)$ in step 1 of the algorithm. Substitution in condition (6.2) gives

$$\begin{aligned}
&\psi_{\rm ub}(t^+, \widehat{x}, \widehat{W}, \widehat{Z}) \\
&\quad = t^+(\mathbf{Tr} G(x^\star(t))W^\star(t) + \mathbf{Tr} F(x^\star(t))Z^\star(t) - \log\det G(x^\star(t))W^\star(t) - l) \\
&\qquad - \log\det F(x^\star(t))Z^\star(t) - n(1 + \log t^+) \\
&\quad = n(t^+/t - 1 - \log(t^+/t)) = \gamma, \qquad (6.5)
\end{aligned}$$

which is a simple nonlinear equation in one variable, with a unique solution $t^+ > t$. We call this variant of the algorithm the *fixed-reduction* algorithm because it uses the same value of t^+/t — and hence achieves a fixed duality gap reduction factor — in each outer iteration. The outline of the fixed-reduction algorithm is as follows.

Fixed-reduction algorithm

given $\gamma > 0$, $t \geq 1$, $x := x^\star(t)$
Find α such that $n(\alpha - 1 - \log\alpha) = \gamma$
repeat
1. $t^+ := \alpha t$
2. Compute $x^\star(t^+)$ starting at x, using the Newton algorithm of section 5
3. $t := t^+$, $x := x^\star(t^+)$

until $n/t \leq \epsilon$

We can be brief in the convergence analysis of the method. Each outer iteration reduces the duality gap by a factor α, so the number of outer iterations is *exactly*

$$\left\lceil \frac{\log(\epsilon^{(0)}/\epsilon)}{\log\alpha} \right\rceil.$$

The inequality (6.3), which was used in the complexity analysis of the previous section, follows from the fact that for $y \geq 1$

$$n(y - 1 - \log y) \leq \frac{n}{2}(y-1)^2,$$

and hence $\alpha \geq 1 + \sqrt{2\gamma/n}$.

This convergence analysis also reveals the limitation of the fixed reduction method: the number of outer iterations is never better than the number predicted by the theoretical analysis. The upper bound on the total number of Newton iterations (6.4) is also a good estimate in practice, provided we replace the constant $5 + 11\gamma$ with an empirically determined estimate such as $3 + 0.7\gamma$ (see Figure 5.1). The purpose of the next section is to develop a method with the same worst-case complexity as the fixed-reduction algorithm but a much better performance in practice.

6.3. Primal-dual long-step algorithm. It is possible to use much larger values of t^+/t, and hence achieve larger gap reduction per outer iteration, by using a better choice for $\widehat{x}$, $\widehat{W}$, and $\widehat{Z}$ in Step 1 of the path-following algorithm.

A natural choice for $\widehat{x}$ is to take a point along the tangent to the central path, i.e.,

$$\widehat{x} = x^\star(t) + p\frac{\partial x^\star(t)}{\partial t},$$

for some $p > 0$, where the tangent direction is given by (4.6). Substitution in (6.2) gives a nonlinear equation from which t^+ and p can be determined. Taking the idea one step further, one can allow $\widehat{W}$ and $\widehat{Z}$ to vary along the tangent to the dual central path, i.e., take

$$\widehat{W} = W^\star(t) + q\frac{\partial W^\star(t)}{\partial t}, \qquad \widehat{Z} = Z^\star(t) + q\frac{\partial Z^\star(t)}{\partial t}$$

for some $q > 0$, with the tangent directions given by (4.7) and (4.8). Equation (6.2) then has three unknowns: t^+, the primal step length p, and the dual step length q. The fixed-reduction update of the previous section uses the solution $t^+ = \alpha t$, $p = q = 0$; an efficient method for finding a solution with larger t^+ is described below.

The outline of the long-step algorithm is as follows.

Primal-dual long-step algorithm

given $\gamma > 0$, $t \geq 1$, $x := x^\star(t)$, $W := W^\star(t)$, $Z := Z^\star(t)$
Find α such that $n(\alpha - 1 - \log\alpha) = \gamma$
repeat

1. *Compute tangent to central path.* $\delta x := \frac{\partial x^\star(t)}{\partial t}$, $\delta W := \frac{\partial W^\star(t)}{\partial t}$, $\delta Z := \frac{\partial Z^\star(t)}{\partial t}$
2. *Parameter selection and predictor step.*
   2a. $t^+ := \alpha t$
   repeat {
      2b. $\widehat{p}, \widehat{q} = \operatorname{argmin}_{p,q} \psi_{ub}(t^+, x + p\delta x, W + q\delta W, Z + q\delta Z)$
      2c. Compute t^+ from $\psi_{\rm ub}(t^+, x + \widehat{p}\delta x, W + \widehat{q}\delta W, Z + \widehat{q}\delta Z) = \gamma$
   }
   2d. $\widehat{x} = x + \widehat{p}\delta x$
3. *Centering step.* Compute $x^\star(t^+)$ starting at $\widehat{x}$, using the Newton algorithm of section 5
4. *Update.* $t := t^+$, $x := x^\star(t^+)$, $W := W^\star(t^+)$, $Z := Z^\star(t^+)$

until $n/t \leq \epsilon$

Again we assume exact centering in step 3. In practice, approximate minimization works, provided one includes a small correction to the formulas of the tangent directions; see [70].

Step 2 computes a solution to (6.2), using a technique illustrated in Figure 6.1. The figure shows four iterations of the inner loop of step 2 (for an instance of the problem family described in section 9). With a slight abuse of notation, we write $\psi_{\rm ub}(t^+, p, q)$ instead of

$$\psi_{\rm ub}(t^+, x^\star(t) + p\delta x, W^\star(t) + q\delta W, Z^\star(t) + q\delta Z). \tag{6.6}$$

We start at the value $t^{(0)} = t$, at the left end of the horizontal axis. The first curve (marked $\psi_{\rm ub}(t^+, 0, 0)$) shows (6.6) as a function of t^+, with $p = q = 0$, which

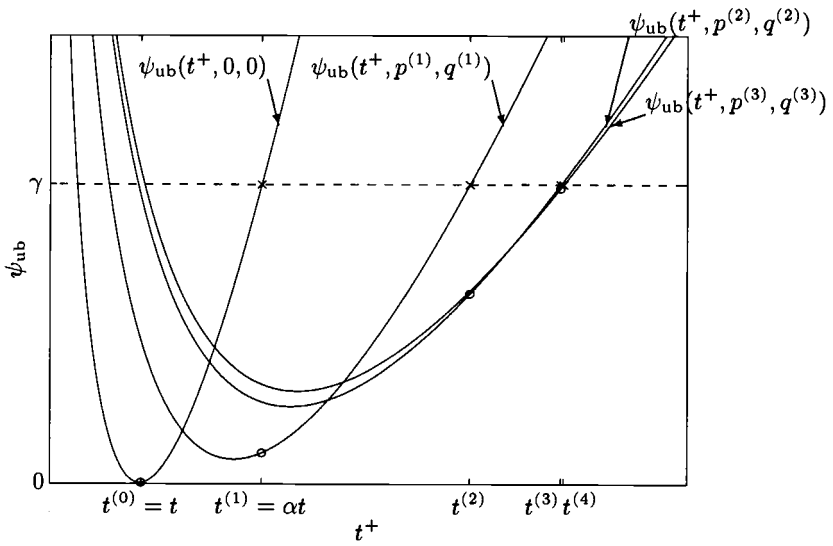


FIG. 6.1. *Parameter selection and predictor step in long-step algorithm alternates between minimizing $\psi_{ub}(t^+, p, q)$ over primal step length p and dual step length q, and then increasing t^+ until $\psi_{ub}(t^+, p, q) = \gamma$.*

simplifies to

$$\psi_{ub}(t^+, x^\star(t), W^\star(t), Z^\star(t)) = n(t^+/t - 1 - \log(t^+/t))$$

(see section 6.2). This function is equal to zero for $t^+ = t$, and equal to γ for the short-step update $t^+ = \alpha t$. We then do the first iteration of the inner loop of step 2. Keeping t^+ fixed at its value $t^{(1)}$, we minimize the function (6.6) over p and q (step 2b). This produces new values $\widehat{p} = p^{(1)}$ and $\widehat{q} = q^{(1)}$ with a value of $\psi_{ub} < \gamma$. This allows us to increase t^+ again (step 2c). The second curve in the figure (labeled $\psi_{ub}(t^+, p^{(1)}, q^{(1)})$) shows the function (6.6) as a function of t^+ with fixed values $p = p^{(1)}$, $q = q^{(1)}$. The intersection with $\psi_{ub} = \gamma$ gives the next value $t^+ = t^{(2)}$.

These two steps (2b, 2c) are repeated either for a fixed number of iterations or until t^+ converges (which in the example of Figure 6.1 happens after four or five iterations). Note that in each step 2c, we increase t^+ so that, in particular, the final value of t^+ will be at least as large as its initial (short-step) value, $t^+ = \alpha t$. Thus, the complexity analysis for the short-step method still applies.

In practice, the inner loop (2b, 2c) often yields a value of t^+ considerably larger than the short-step value αt, while maintaining the same upper bound on the number of Newton steps required to compute the next iterate $x^\star(t^+)$. In the example shown in the figure, the final value of t^+ is about a factor of 2.5 larger than the short-step value; in general, a factor of 10 is not uncommon.

Using some preprocessing we will describe in section 8, the cost of the inner loop (2b, 2c) is very small, in most cases negligible compared with the cost of computing the tangent vectors.

Finally, note that the dual variables Z and W are not used in the fixed-reduction algorithm. In the primal-dual long-step algorithm they are used only in the predictor step to allow a larger step size p.

7. Preliminary phases. The algorithm starts at a central point $x^\star(t)$, for some $t \geq 1$. In this section we discuss how to select the initial t, and how to compute such a point.

Feasibility. If no strictly primal feasible point is known, one has to precede the algorithm with a first phase to solve the (SDP) feasibility problem: find x that satisfies $G(x) \succ 0$, $F(x) \succ 0$. More details can be found in [69].

Choice of initial *t*. We now consider the situation where a strictly primal feasible point $x^{(0)}$ is known, but $x^{(0)}$ is not on the central path. In that case one has to select an appropriate initial value of t and compute a central point by Newton's method starting at $x^{(0)}$. In theory (and often in practice) the simple choice $t = 1$ works.

It is not hard, however, to imagine cases where the choice $t = 1$ would be inefficient in practice. Suppose, for example, that the initial $x^{(0)}$ is very near $x^\star(100)$, so a reasonable initial value of t is 100 (but we don't know this). If we set $t = 1$, we expend many Newton iterations "going backwards" up the central path towards the point $x^\star(1)$. Several outer iterations, and many Newton steps later, we find ourselves back near where we started, around $x^\star(100)$.

If strictly dual feasible points $W^{(0)}$, $Z^{(0)}$ are known, then we start with a known duality gap α associated with $x^{(0)}$, $W^{(0)}$, and $Z^{(0)}$. A very reasonable initial choice for t is then $t = \max\{1, n/\alpha\}$, since when $t = n/\alpha$, the centering stage computes central points with the same duality gap as the initial primal and dual solutions. In particular, the preliminary centering stage does not increase the duality gap (as it would in the scenario sketched above).

We can also interpret and motivate the initial value $t = n/\alpha$ in terms of the function $\psi_{\rm ub}(t, x^{(0)}, W^{(0)}, Z^{(0)})$, which provides an upper bound on the number of Newton steps required to compute $x^\star(t)$ starting at $x^{(0)}$. From the definition (5.5) we have

$$\psi_{\rm ub}(t, x^{(0)}, W^{(0)}, Z^{(0)}) = t\alpha + \log\det F(x^{(0)})^{-1} + \log\det {Z^{(0)}}^{-1} - n(1 + \log t),$$

which shows that the value $t = n/\alpha$ minimizes $\psi_{\rm ub}(t, x^{(0)}, W^{(0)}, Z^{(0)})$. Thus, the value $t = n/\alpha$ is the value which minimizes the upper bound on the number of Newton steps required in the preliminary centering stage.

A heuristic preliminary stage. When no initial dual feasible Z, W (and hence duality gap) are known, choosing an appropriate initial value of t can be difficult. We have had practical success with a variation on Newton's method that adapts the value of t at each step based on the (square of) the Newton decrement $\lambda(x, t)$,

$$\lambda(x,t)^2 = \nabla\varphi_p(t,x)^T \left(\nabla^2\varphi_p(t,x)\right)^{-1} \nabla\varphi_p(t,x),$$

which serves as a measure of proximity to the central path. It is a convex function of t, and is readily minimized in t for fixed x. Our heuristic preliminary phase is as follows.

Preliminary centering phase

given strictly feasible x
$t := 1$
repeat {
1. $t := \max\{1, \operatorname{argmin} \lambda(x,t)\}$
2. $\delta x = -\left(\nabla^2\varphi_p(t,x)\right)^{-1} \nabla\varphi_p(t,x)$

3. $\widehat{h} = \operatorname{argmin} \varphi_p(t, x + h\delta x^N)$
} **until** $\lambda \leq \delta$

Thus, we adjust t each iteration to make the Newton decrement for the current x as small as possible (subject to the condition that t remains greater than one).

8. Efficient line and plane searches. In this section we describe some simple preprocessing that allows us to implement the line search in the Newton method of section 5 and the plane search of section 6.3 very efficiently.

Line search in Newton's method. We first consider the line search in Newton's method of section 5. Let λ_k, $k = 1, \ldots, l$, be the generalized eigenvalues of the pair $\sum_{i=1}^m \delta x_i^N G_i$, $G(x)$, and λ_k, $k = l+1, \ldots, l+n$, be the generalized eigenvalues of the pair $\sum_{i=1}^m \delta x_i^N F_i$, $F(x)$, where δx^N is the Newton direction at x. We can write $\varphi_p(t, x + h\delta x^N)$ in terms of these eigenvalues as

$$f(h) = \varphi_p(t, x + h\delta x^N) = \varphi_p(t, x) + hc^T \delta x^N + t \sum_{k=1}^{l} \log \frac{1}{1 + h\lambda_k} + \sum_{k=l+1}^{l+n} \log \frac{1}{1 + h\lambda_k}.$$

Evaluating the first and second derivatives $f'(h)$, $f''(h)$ of this (convex) function of $h \in \mathbf{R}$ requires only $O(n + l)$ operations (once the generalized eigenvalues λ_i have been computed). In most cases, the cost of the preprocessing, i.e., computing the generalized eigenvalues λ_i, exceeds the cost of minimizing over h but is small compared with the cost of computing the Newton direction. The function $\varphi_p(t, x + h\delta x^N)$ can therefore be efficiently minimized using standard line search techniques.

Plane search in long-step path-following method. A similar idea applies to the plane search of section 6.3. In step 2c of the primal-dual long-step algorithm we minimize the function $\psi_{\mathrm{ub}}(t, x + p\delta x, W + q\delta W, Z + q\delta Z)$ over p and q, where δx, δW, δZ are tangent directions to the central path. We can again reduce the function to a convenient form

$$\begin{aligned} &\psi_{\mathrm{ub}}(t, x + p\delta x, W + q\delta W, Z + q\delta Z) \\ &\quad = \psi_{\mathrm{ub}}(t, x, W, Z) + p\beta_1 + q\beta_2 + t \sum_{k=1}^{l} \log \frac{1}{1 + p\lambda_k} + \sum_{k=l+1}^{l+n} \log \frac{1}{1 + p\lambda_k} \\ &\qquad + t \sum_{k=1}^{l} \log \frac{1}{1 + q\mu_k} + \sum_{k=l+1}^{l+n} \log \frac{1}{1 + q\mu_k}, \end{aligned} \tag{8.1}$$

where λ_k, $k = 1, \ldots, l$, are the generalized eigenvalues of the pair $\sum_{i=1}^m \delta x_i G_i$, $G(x)$ and λ_k, $k = l+1, \ldots, l+n$, are the generalized eigenvalues of the pair $\sum_{i=1}^m \delta x_i F_i$, $F(x)$; μ_k, $k = 1, \ldots, l$, are the generalized eigenvalues of the pair δW, W, and μ_k, $k = l+1, \ldots, l+n$, are the generalized eigenvalues of the pair δZ, Z. The coefficients β_1 and β_2 are

$$\beta_1 = c^T \delta x, \quad \beta_2 = \mathbf{Tr} G_0 \delta W + \mathbf{Tr} F_0 \delta Z.$$

The first and second derivatives of the function (8.1) with respect to p and q can again be computed at a low cost of $O(l + n)$, and therefore the minimum of ψ_{ub} over the plane can be determined very cheaply, once the generalized eigenvalues have been computed.

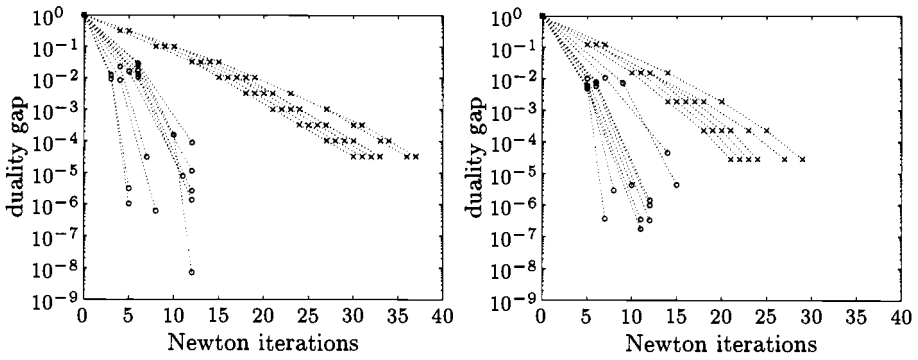


FIG. 9.1. *Duality gap versus number of Newton steps for randomly generated max-det problems of dimension $l = 10$, $n = 10$, $m = 10$. Left: $\gamma = 10$. Right: $\gamma = 50$. The crosses are the results for the fixed-reduction method; the circles are the results for the long-step method. Every cross/circle represents the gap at the end of an outer iteration.*

In summary, the cost of line or plane search is basically the cost of preprocessing (computing certain generalized eigenvalues), which is usually negligible compared to the rest of algorithm (e.g., determining a Newton or tangent direction).

One implication of efficient line and plane searches is that the total number of Newton steps serves as a good measure of the overall computing effort.

9. Numerical examples.

Typical convergence. The first experiment (Figure 9.1) compares the convergence of the fixed-reduction method and the long-step method. The left-hand plot shows the convergence of both methods for $\gamma = 10$; the right-hand plot shows the convergence for $\gamma = 50$. Duality gap is shown vertically on a logarithmic scale ranging from 10^0 at the top to 10^{-9} at the bottom; the horizontal axis is the total number of Newton steps. Each outer iteration is shown as a symbol on the plot ("$\circ$" for the long-step and "$\times$" for the short-step method). Thus, the horizontal distance between two consecutive symbols shows directly the number of Newton steps required for that particular outer iteration; the vertical distance shows directly the duality gap reduction factor t^+/t.

Problem instances were generated as follows: $G_0 \in \mathbf{R}^{l\times l}$, $F_0 \in \mathbf{R}^{n\times n}$ were chosen random positive definite (constructed as U^TU with the elements of U drawn from a normal distribution $\mathcal{N}(0,1)$); the matrices G_i, F_i, $i = 1,\ldots,m$, were random symmetric matrices, with elements drawn from $\mathcal{N}(0,1)$; $c_i = \mathbf{Tr}G_i + \mathbf{Tr}F_i$, $i = 1,\ldots,m$. This procedure ensures that the problem is primal and dual feasible ($x = 0$ is primal feasible; $Z = I$, $W = I$ is dual feasible), and hence bounded below. We start on the central path, with initial duality gap one.

We can make the following observations.

- The convergence is very similar over all problem instances. The number of iterations required to reduce the duality gap by a factor 1000 ranges between 5 and 50. As expected, the long-step method performs much better than the fixed-reduction method, and typically converges in less than 15 iterations.
- The fixed-reduction method converges almost linearly. The duality gap reduction t^+/t per outer iteration can be computed from equation (6.5): $t^+/t = 3.14$ for $\gamma = 10$, and $t^+/t = 8.0$ for $\gamma = 50$. The number of Newton iterations per outer iteration is less than five in almost all cases, which is much less than

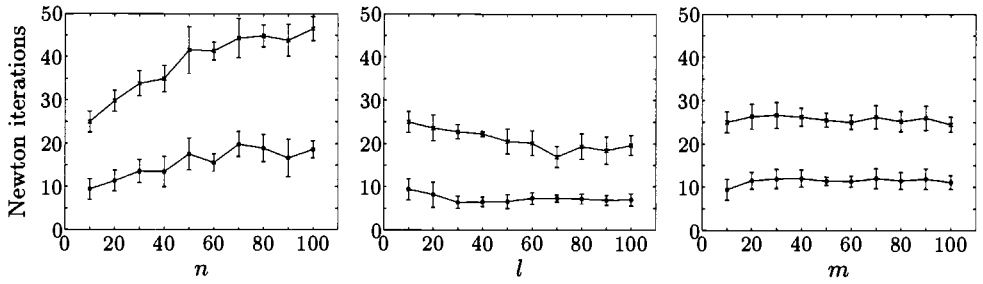


FIG. 9.2. *Newton iterations versus problem size for family of random problems. Fixed-reduction method (top curve) and long-step method (lower curve).* $\gamma = 10$. *Left:* $l = 10$, $n = 10$–100, $m = 10$. *Middle:* $l = 10$–100, $n = 10$, $m = 10$. *Right:* $l = 10$, $n = 10$, $m = 10$–100. *The curves give the average over* 10 *problem instances. The error bars indicate the standard deviation.*

the upper bound $5 + 11\gamma$. (Remember that this bound is a combination of two conservative estimates: Theorem 5.1 is conservative; see Figure 5.1. In addition we have replaced (6.1) with the weaker condition (6.2).)

- The long-step method takes a few more Newton iterations per centering step but achieves a much larger duality gap reduction. Moreover, the convergence accelerates near the optimum.
- Increasing γ has a large effect on the fixed-reduction method but only little effect on the long-step method.

Complexity versus problem size. Figure 9.2 shows the influence of the problem dimension on the convergence. For each triplet (m, n, l) we generated 10 problem instances as above. We plot the number of Newton iterations to reduce the duality gap by a factor 1000, starting with duality gap 1. The plot shows the average number of Newton steps and the standard deviation. The top curve shows the results for the fixed-reduction method, the lower curve is for the long-step method.

- The number of Newton iterations in the short-step method depends on n as $O(\sqrt{n})$. This is easily explained from the convergence analysis of section 6.2: We have seen that the number of outer iterations grows as $\sqrt{n}$, in theory and in practice, and hence the practical behavior of the fixed-reduction method is very close to the worst-case behavior.
- We see that the number of iterations for the long-step method lies between 5 and 20, and is very weakly dependent on problem size.

Figure 9.3 shows similar results for a family of experiment design problems (2.9) in $\mathbf{R}^{10}$, including a 90-10 constraint (2.10). The points $v^{(i)}$, $i = 1, \ldots, M$, were generated from a normal distribution $\mathcal{N}(0, I)$ on $\mathbf{R}^p$. Note that the dimensions of the corresponding max-det problem are $m = 2M$, $n = 3M + 1$, $l = p$. Figure 9.3 confirms the conclusions of the previous experiment: it shows that the complexity of the fixed-reduction method grows as $\sqrt{n}$, while the complexity of the long-step method is almost independent of problem size.

10. Conclusion. The max-det problem (1.1) is a (quite specific) convex extension of the semidefinite programming problem, and hence includes a wide variety of convex optimization problems as special cases. Perhaps more importantly, max-det problems arise naturally in many areas, including computational geometry, linear algebra, experiment design, linear estimation, and information and communication theory. We have described several of these applications.

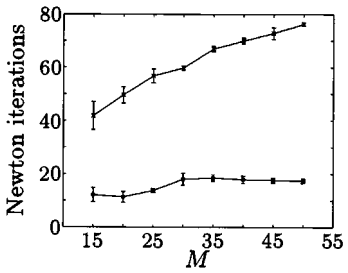


FIG. 9.3. *Newton iterations versus problem size for family of experiment design problems of section* 2.4 *including* 90-10 *rule. Fixed-reduction method (top curve) and long-step method (lower curve).* $\gamma = 10$, $p = 10$, $M = 15$–50. *The curves show the average over* 10 *problem instances. The error bars indicate the standard deviation.*

Some of the applications have been studied extensively in the literature, and in some cases analytic solutions or efficient specialized algorithms have been developed. We have presented an interior-point algorithm that solves *general* max-det problems efficiently. The method can be applied to solve max-det problems for which no specialized algorithm is known; in cases where such a method exists, it opens the possibility of adding useful LMI constraints, which is an important advantage in practice.

We have proved a worst-case complexity of $O(\sqrt{n})$ Newton iterations. Numerical experiments indicate that the behavior is much better in practice: the method typically requires a number of iterations that lies between 5 and 50, almost independent of problem dimension. The total computational effort is therefore determined by the amount of work per iteration, i.e., the computation of the Newton directions, and therefore depends heavily on the problem structure. When no structure is exploited, the Newton directions can be computed from the least squares formulas in [70], which require $O((n^2 + l^2)m^2)$ operations, but important savings are possible whenever we specialize the general method of this paper to a specific problem class.

Acknowledgments. We have many people to thank for useful comments, suggestions, and pointers to references and applications (including some applications that did not make it into the final version). In particular, we thank Paul Algoet, Kurt Anstreicher, Eric Beran, Patrick Dewilde, Valerii Fedorov, Gene Golub, Bijit Halder, Bill Helton, Istvan Kollar, Lennart Ljung, Tomas MacKelvey, Erik Ordentlich, Art Owen, Randy Tobias, Rick Wesel, and Henry Wolkowicz. We also thank the two reviewers and Michael Overton for very useful comments.

REFERENCES

[1] F. ALIZADEH, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim., 5 (1995), pp. 13–51.

[2] T. W. ANDERSON, *Statistical inference for covariance matrices with linear structure*, in Multivariate Analysis II. Proceedings of the Second International Symposium on Multivariate Analysis, P. R. Krishnaiah, ed., Academic Press, New York, 1969, pp. 55–66.

[3] T. W. ANDERSON, *Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices*, in Essays in Probability and Statistics, R. C. Bose et al., eds., University of North Carolina Press, Chapel Hill, NC, 1970, pp. 1–24.

[4] K. M. ANSTREICHER AND M. FAMPA, *A Long-Step Path Following Algorithm for Semidefinite Programming Problems*, Tech. report, Department of Management Sciences, University of Iowa, Iowa City, IA, 1996.

[5] K. M. Anstreicher, M. Fampa, J. Lee, and J. Williams, *Continuous relaxations for constrained maximum-entropy sampling*, in Integer Programming and Combinatorial Optimization, W. H. Cunningham, S. T. McCormick, and M. Queyranne, eds., Lecture Notes in Comput. Sci. 1084, Springer-Verlag, New York, 1996, pp. 234–248.

[6] J. T. Aslanis and J. M. Cioffi, *Achievable information rates on digital subscriber loops: limiting information rates with crosstalk noise*, IEEE Trans. Comm., 40 (1992), pp. 361–372.

[7] A. C. Atkinson and A. N. Donev, *Optimum experiment designs*, Oxford Statist. Sci. Ser., Oxford University Press, London, 1992.

[8] M. Bakonyi and H. J. Woerdeman, *Maximum entropy elements in the intersection of an affine space and the cone of positive definite matrices*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 360–376.

[9] E. R. Barnes, *An algorithm for separating patterns by ellipsoids*, IBM J. Res. Develop., 26 (1982), pp. 759–764.

[10] W. W. Barrett, C. R. Johnson, and M. Lundquist, *Determinantal formulation for matrix completions associated with chordal graphs*, Linear Algebra Appl., 121 (1989), pp. 265–289.

[11] D. Bertsekas and I. Rhodes, *Recursive state estimation for a set-membership description of uncertainty*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 117–128.

[12] S. Boyd and L. El Ghaoui, *Method of centers for minimizing generalized eigenvalues*, Linear Algebra Appl., 188 (1993), pp. 63–111.

[13] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*, Stud. Appl. Math. 15, SIAM, Philadelphia, PA, 1994.

[14] S. Boyd and L. Vandenberghe, *Introduction to convex optimization with engineering applications*, Lecture Notes, Information Systems Laboratory, Stanford University, Stanford, CA, 1995; available online from http://www-isl.stanford.edu/people/boyd/392/ee392x.html.

[15] J. P. Burg, D. G. Luenberger, and D. L. Wenger, *Estimation of structured covariance matrices*, Proc. IEEE, 30 (1982), pp. 963–974.

[16] F. L. Chernousko, *Guaranteed estimates of undetermined quantities by means of ellipsoids*, Soviet Math. Dokl., 21 (1980), pp. 396–399.

[17] F. L. Chernousko, *State Estimation for Dynamic Systems*, CRC Press, Boca Raton, FL, 1994.

[18] M. Cheung, S. Yurkovich, and K. M. Passino, *An optimal volume ellipsoid algorithm for parameter estimation*, IEEE Trans. Automat. Control, AC-38 (1993), pp. 1292–1296.

[19] D. Cook and V. Fedorov, *Constrained optimization of experimental design*, Statistics, 26 (1995), pp. 129–178.

[20] T. M. Cover and S. Pombra, *Gaussian feedback capacity*, IEEE Trans. Inform. Theory, 35 (1989), pp. 37–43.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, 1991.

[22] C. Davis, W. M. Kahan, and H. F. Weinberger, *Norm-preserving dilations and their applications to optimal error bounds*, SIAM J. Numer. Anal., 19 (1982), pp. 445–469.

[23] J. Deller, *Set membership identification in digital signal processing*, IEEE Trans. Acoustics, Speech and Signal Processing, 6 (1989), pp. 4–20.

[24] J. R. Deller, M. Nayeri, and S. F. Odeh, *Least-square identification with error bounds for real-time signal processing and control*, Proc. IEEE, 81 (1993), pp. 815–849.

[25] A. Dembo, *The relation between maximum likelihood estimation of structured covariance matrices and periodograms*, IEEE Trans. Acoustics, Speech and Signal Processing, 34 (1986), pp. 1661–1662.

[26] A. Dembo, C. L. Mallows, and L. A. Shepp, *Embedding nonnegative definite Toeplitz matrices in nonnegative definite circulant matrices, with application to covariance estimation*, IEEE Trans. Inform. Theory, 35 (1989), pp. 1206–1212.

[27] A. P. Dempster, *Covariance selection*, Biometrics, 28 (1972), pp. 157–175.

[28] D. den Hertog, *Interior Point Approach to Linear, Quadratic and Convex Programming*, Kluwer Academic Publishers, Norwell, MA, 1993.

[29] P. Dewilde and E. F. A. Deprettere, *The generalized Schur algorithm: approximations and hierarchy*, in Topics in Operator Theory and Interpolation, I. Gohberg, ed., Birkhaüser-Verlag, Basel, Switzerland, 1988, pp. 97–116.

[30] P. Dewilde and Z. Q. Ning, *Models for Large Integrated Circuits*, Kluwer Academic Publishers, Norwell, MA, 1990.

[31] H. Dym and I. Gohberg, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[32] L. El Ghaoui, *Robustness analysis of semidefinite programs and applications to matrix completion problems*, presented at MTNS-96, St. Louis, MO, 1996.
[33] V. V. Fedorov, *Theory of Optimal Experiments*, Academic Press, New York 1971.
[34] A. Fiacco and G. McCormick, *Nonlinear programming: sequential unconstrained minimization techniques*, Wiley, 1968; SIAM Philadelphia, PA, 1990.
[35] E. Fogel, *System identification via membership set constraints with energy constrained noise*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 752–757.
[36] E. Fogel and Y. Huang, *On the value of information in system identification-bounded noise case*, Automatica, 18 (1982), pp. 229–238.
[37] 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.
[38] J. W. Helton and H. J. Woerdeman, *Symmetric Hankel operators: Minimal norm extensions and eigenstructures*, Linear Algebra Appl., 185 (1993), pp. 1–19.
[39] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Grundlehren der mathematischen Wissenschaften 305, Springer-Verlag, New York, 1993.
[40] G. B. Jávorzky, I. Kollár, L. Vandenberghe, S. Boyd, and S.-P. Wu, *Optimal excitation signal design for frequency domain system identification using semidefinite programming*, in Proc. 8th IMEKO TC4 Symposium on Recent Advances in Electrical Measurements, Budapest, Hungary, 1996, pp. 192–197.
[41] C. R. Johnson, *Positive definite completions: a guide to selected literature*, in Signal Processing. Part I: Signal Processing Theory, L. Auslander, T. Kailath, and S. Mitter, eds., IMA Vol. Math. Appl., Springer, New York, 1990, pp. 169–188.
[42] C. R. Johnson, B. Kroschel, and H. Wolkowicz, *An Interior-Point Method for Approximate Positive Semidefinite Completions*, Tech. report CORR Report 95-11, University of Waterloo, Waterloo, ON, Canada, 1995; Comput. Optim. Appl., to appear.
[43] L. G. Khachiyan and M. J. Todd, *On the complexity of approximating the maximal inscribed ellipsoid for a polytope*, Math. Programming, 61 (1993), pp. 137–159.
[44] C. W. Ko, J. Lee, and K. Wayne, *A Spectral Bound for D-optimality*, Tech. report, Department of Mathematics, University of Kentucky, Lexington, KY, 1996.
[45] J. Lee, *Constrained Maximum Entropy Sampling*, Tech. report, Department of Mathematics, University of Kentucky, Lexington, KY, 1995; Oper. Res., to appear.
[46] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177.
[47] A. S. Lewis and M. L. Overton, *Eigenvalue optimization*, Acta Numerica, 5 (1996), pp. 149–190.
[48] D. G. Luenberger, *Optimization By Vector Space Methods*, John Wiley, New York, 1969.
[49] M. E. Lundquist and C. R. Johnson, *Linearly constrained positive definite completions*, Linear Algebra Appl., 150 (1991), pp. 195–207.
[50] G. Naevdal and H. J. Woerdeman, *Partial matrix contractions and intersections of matrix balls*, Linear Algebra Appl., 175 (1992), pp. 225–238.
[51] Y. Nesterov and A. Nemirovsky, *Interior-Point Polynomial Methods in Convex Programming*, Studies Appl. Math. 13, SIAM, Philadelphia, PA, 1994.
[52] J. P. Norton, *An Introduction to Identification*, Academic Press, New York, 1986.
[53] J. P. Norton, *Identification and application of bounded-parameter models*, Automatica, 23 (1987), pp. 497–507.
[54] L. Pronzato and E. Walter, *Minimum-volume ellipsoids containing compact sets: Application to parameter bounding*, Automatica, 30 (1994), pp. 1731–1739.
[55] F. Pukelsheim, *Optimal Design of Experiments*, Wiley, 1993.
[56] J. Renegar, *A polynomial-time algorithm, based on Newton's method, for linear programming*, Math. Programming, 40 (1988), pp. 59–93.
[57] R. T. Rockafellar, *Convex Analysis*, 2nd ed., Princeton University Press, Princeton, NJ, 1970.
[58] J. B. Rosen, *Pattern separation by convex programming*, J. Math. Anal. Appl., 10 (1965), pp. 123–134.
[59] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley, 1987.
[60] S. S. Sapatnekar, *A Convex Programming Approach to Problems in VLSI Design*, Ph.D. thesis, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1992.
[61] L. L. Scharf, *Statistical Signal Processing*, Addison–Wesley, Reading, MA, 1991.
[62] F. C. Schweppe, *Recursive state estimation: Unknown but bounded errors and system inputs*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 22–28.
[63] F. C. Schweppe, *Uncertain Dynamic Systems*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[64] G. Sonnevend, *An 'analytical centre' for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming*, in Lecture Notes in Control and Inform. Sci. 84, Springer-Verlag, New York, 1986, pp. 866–878.

[65] G. Sonnevend, *Applications of analytic centers*, in Numerical Linear Algebra, Digital Signal Processing, and Parallel Algorithms, G. Golub and P. V. Dooren, eds., NATO ASI Series F70, Springer-Verlag, New York, 1991, pp. 617–632.

[66] S. Tarasov, L. Khachiyan, and I. Erlikh, *The method of inscribed ellipsoids*, Soviet Math. Dokl., 37 (1988), pp. 226–230.

[67] D. M. Titterington, *Optimal design: some geometric aspects of D-optimality*, Biometrika, 62 (1975), pp. 313–320.

[68] L. Vandenberghe and S. Boyd, *Connections between semi-infinite and semidefinite programming*, in Proc. International Workshop on Semi-Infinite Programming, R. Reemtsen and J.-J. Rueckmann, eds., Kluwer Academic Publishers, Norwell, MA, 1996, to appear.

[69] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

[70] L. Vandenberghe, S. Boyd, and S.-P. Wu, *Determinant Maximization with Linear Matrix Inequality Constraints*, Tech. report, Information Systems Laboratory, Stanford University, Stanford, CA, 1996.

[71] E. Walter and H. Piet-Lahanier, *Estimation of parameter bounds from bounded-error data: a survey*, Math. Comput. Simulation, 32 (1990), pp. 449–468.

[72] P. Whittle, *Optimization over Time. Dynamic Programming and Stochastic Control*, John Wiley, New York, 1982.

[73] A. Wilhelm, *Computing Optimal Designs by Bundle Trust Methods*, Tech. report 314, Institüt für Mathematik, Universität Augsburg, 1994.

[74] A. Wilhelm, *Algorithms for differentiable and nondifferentiable experimental design problems*, in SoftStat '95–Advances in Statistical Software, F. Faulbaum and W. Bandilla, eds., Lucius and Lucius, Stuttgart, 1995, pp. 527–534.

[75] H. S. Witsenhausen, *Sets of possible states of linear systems given perturbed observations*, IEEE Trans. Automat. Control, 13 (1968), pp. 556–558.

[76] S.-P. Wu, L. Vandenberghe, and S. Boyd, *maxdet: Software for Determinant Maximization Problems. User's Guide, Alpha Version*, Stanford University, Stanford, CA, 1996.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 534–540, April 1998

015

OVERCOMING INSTABILITY IN COMPUTING THE FUNDAMENTAL MATRIX FOR A MARKOV CHAIN*

DANIEL P. HEYMAN† AND DIANNE P. O'LEARY‡

Abstract. We present an algorithm for solving linear systems involving the probability or rate matrix for a Markov chain. It is based on a UL factorization but works only with a submatrix of the factor U. We demonstrate its utility on Erlang-B models as well as more complicated models of a telephone multiplexing system.

Key words. Markov chains, fundamental matrix, decision process

AMS subject classifications. 65F05, 62M05

PII. S0895479896301753

1. Introduction. Markov chain models can lend insight into the behavior of many physical systems, such as telephone networks, highway systems, and ATM switching networks. These models are based on properties of a matrix P whose entries depend on the probabilities of transition from one state to another, or on the arrival and departure rates for customers. The matrix P is nonnegative. If we define D to be a diagonal matrix whose diagonal entries are the rowsums for P, then the matrix $D - P$ has zero rowsums. In other words

$$(D - P)e = 0,$$

where e is the column vector of all ones. Thus, $D - P$ has a zero eigenvalue, and we denote its left eigenvector, normalized so that its entries sum to one, as the row vector π^T:

$$\pi^T(D - P) = 0^T, \qquad \pi^T e = 1\,.$$

The vector π gives information about the long-term behavior of the system; for example, if the entries in P are transition probabilities (so that $D = I$), then π is the stationary vector for the chain.

Systems analysts are interested in other computational quantities that give information about the short-term behavior of the chain. The *fundamental matrix* is defined to be

$$F = (D - P - e\pi^T)^{-1},$$

and the *group generalized inverse* of $A = D - P$ is

$$A^{\#} \equiv F - e\pi^T.$$

(See, for example, [6].) The entries in these matrices are useful in computing mean first passage times, in computing biases in the entries in π as approximations to

*Received by the editors April 10, 1996; accepted for publication (in revised form) by D. Calvetti April 10, 1997.

http://www.siam.org/journals/simax/19-2/30175.html

†AT&T Labs, 101 Crawfords Corner Road, Holmdel, NJ 07733 (Daniel.Heyman@att.com).

‡Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (oleary@cs.umd.edu). The research of this author was supported by National Science Foundation grant CCR-95-03126.

expected number of visits, and in determining decision rules to govern the control of the system. See [2] and [5] for some discussion of these applications.

The Grassmann, Taksar, and Heyman (GTH) algorithm [1] is an efficient algorithm for determining a factorization of the matrix $D - P$. From this factorization, all the other quantities can easily be computed.

The GTH algorithm can be interpreted as a variant of Gauss elimination that differs from the usual LU form in two ways:

1. The elimination proceeds from the bottom of the matrix to the top (rather than top-to-bottom) and thus produces factors

$$D - P = UL,$$

   where U is an upper triangular matrix with diagonal elements equal to -1, and L is a lower triangular matrix, with top row zero.
2. Since the rowsums of $D - P$ are zero, so are the rowsums of the matrix L. We compute the main diagonal elements of L to satisfy this constraint, rather than using the usual Gaussian elimination formulas. This modification has been shown to provide a very strong form of numerical stability when making use of these factors to compute the stationary vectors [7].

We assume that the rows of the matrix (one for each state of the chain) are numbered $\{0, 1, \ldots, n\}$, we let `P(i:j,k:l)` denote the submatrix of P consisting of elements in rows $i, i+1, \ldots, j$ and columns $k, k+1, \ldots, l$, and we let `sum(P(i,j:k))` denote the sum of the elements in row i and columns j through k. Then the GTH algorithm computes $UL = D - P$. It can be done using no additional matrix or vector storage as follows:

```
For i=n, n-1, ... , 1,
   s=sum(P(i,0:i-1))
   P(i,i)=-s
   P(0:i-1,i)=P(0:i-1,i)/s
   P(0:i-1,0:i-1)=P(0:i-1,0:i-1) + P(0:i-1,i)*P(i,0:i-1)
end for
P(0,0)=0
```

Then the matrix L is stored in the lower triangular part of P (including the main diagonal), and U is in the upper triangular, with main diagonal elements understood to be equal to -1. The entire factorization process takes $2/3\, n^3 + O(n^2)$ operations.

Once the UL factors of $D - P$ are determined, it is easy to compute the stationary vector from

$$\pi^T U = z,$$

where z is the first row of the identity matrix. Then the fundamental matrix can be computed from

$$ULF = I - e\pi^T,$$

with normalization $\pi^T F = \pi^T$, or the group generalized inverse from

$$ULA^{\#} = I - e\pi^T$$

with normalization $\pi^T A^{\#} = 0^T$.

TABLE 2.1
Algorithm results on the examples of section 2.

n	r_{orig}	$r_{improved}$	cond(U)	cond($D-P$)
5	4.2e-15	1.0e-15	2.4e+02	8.1e+00
10	3.6e-13	3.0e-15	5.4e+04	2.0e+01
15	1.3e-12	9.0e-15	1.1e+07	3.4e+01
20	4.4e-09	1.5e-14	1.9e+09	5.0e+01
25	1.3e-06	1.7e-14	3.3e+11	6.6e+01
30	2.5e-04	2.9e-14	5.6e+13	8.2e+01
35	2.8e-02	3.2e-14	9.2e+15	9.9e+01
40	2.5e+00	3.3e-14	1.5e+18	1.2e+02
45	4.1e+00	3.8e-14	Inf	1.3e+02
50	1.1e+05	5.5e-14	Inf	1.5e+02

Clearly, the lower triangular matrix L is singular, since its first row is zero. Conventional wisdom says that the matrix U is usually well conditioned, but occasionally this fails to be true, even if the nonzero singular values of $D-P$ are well behaved.

The purpose of this note is to exhibit examples of this phenomenon and to propose a more stable way to use the UL factors to compute the fundamental matrix and other quantities.

2. A troublesome family of examples. Consider an *Erlang-B model* of telephone traffic. Calls arrive as a Poisson process at rate λ to be served by n parallel servers at unit rate. A call that finds all servers busy is discarded.

This model yields a continuous-time Markov chain with states $\{0, 1, \ldots, n\}$. Let p_{ij} be the rate of passage from state i to state j. Then the only nonzero rates are

$$p_{i,i+1} = \lambda, \quad \text{for } 0 \le i < n,$$
$$p_{i,i-1} = i, \quad \text{for } 0 < i \le n.$$

Let the matrix $P = (p_{ij})$ and let D be the diagonal matrix whose entries are the rowsums of P.

Then it is easy to compute the UL factors of $D-P$: both U and L are bidiagonal matrices with nonzero entries

$$u_{ii} = -1, \quad i = 0, \ldots, n,$$
$$u_{i,i+1} = \frac{\lambda}{i+1}, \quad i = 0, \ldots, n-1,$$
$$l_{ii} = -i, \quad i = 0, \ldots, n,$$
$$l_{i,i-1} = i, \quad i = 1, \ldots, n.$$

The unnormalized steady-state probabilities are

$$\pi_i = \frac{\lambda^i}{i!}, \qquad i = 0, \ldots, n.^1$$

If we set $\lambda = n$ and use this data to compute the last column of the fundamental matrix, we get the results in Table 2.1.

[1] Pete Stewart observes that this growth in the elements of the stationary vector can only occur if there is ill conditioning in one of the UL factors. Thus, it may be worthwhile in practice to reorder the states so that the stationary probabilities are decreasing.

These results were computed using double-precision IEEE arithmetic (approximately 16 decimal digits) in MATLAB.

The column labeled r_{orig} gives the norm of the residual vector when the last column of F is computed using the UL factors, i.e.,

$$r_{orig} = \|e_n - \pi_n e - (D - P)f_{orig}\|,$$

where e_n is the last column of the identity matrix and f_{orig} is the result of using forward and back substitution on the linear system

$$ULf_{orig} = e_n - \pi_n e.$$

We see that r_{orig} grows rapidly as n grows, although in exact arithmetic r_{orig} would be zero.

Such large residuals are a symptom of ill conditioning or instability, so the table also gives the condition number of U and the condition number of $D - P$. Here we define the condition number as the ratio of the largest to the smallest singular value of the matrix, although, since $D - P$ is singular, we leave out its zero singular value in computing this ratio. Clearly, the matrix U is rapidly approaching singularity, and thus when we use U to solve for the last column of F, accuracy can be lost.

In the next section we describe an improved algorithm that produces the residuals labeled $r_{improved}$ in the table.

3. A more stable way to use the UL factors. To see what went wrong, we need to look at the null spaces of our various matrices.

Suppose we are solving $(D - P)z = b$, where b is in the range of $D - P$. Then the solution vector z satisfies

$$Lz = y,$$

where y is the solution to

$$Uy = b.$$

Since the top row of L is zero, the top element of y must also be zero in order for the system to have a solution. Thus, before we begin the back substitution on U, we already know the top component of y.

If, due to round-off error and ill conditioning of U, the top component of the computed y fails to be close to zero, then our computation will not produce a good solution.

This insight also leads to a remedy. Instead of solving $Uy = b$, we can solve a system that involves only the last n components of y, knowing that the top one is zero. If we let $\bar{U}$ be the matrix formed by deleting the zeroth column of U, and let $\bar{y}$ be the vector formed by deleting the zeroth element of y, then we can compute $\bar{y}$ by solving the linear system

$$\bar{U}\bar{y} = b.$$

The matrix $\bar{U}$ is not upper triangular (in fact, it is zero *above* the main diagonal for the examples in section 2, since U is bidiagonal). But $\bar{U}$ is always *upper Hessenberg*, with zeros below the first subdiagonal. A sequence of n row operations reduces it to upper triangular form, at a cost of at most $O(n^2)$ floating point operations. Since the

system is compatible, the same sequence of operations reduces the last component of b to zero, permitting back substitution starting with equation $n-1$.

We choose to reduce the matrix $\bar{U}$ to upper triangular form using the LU algorithm with partial pivoting. Just as in the GTH algorithm, only additions and divisions are being performed and no cancellation can occur, and the factorization can be done in-place, except for an auxiliary integer vector of permutation indices. Assume that we apply the LU algorithm to obtain the factorization $\bar{U} = \tilde{L}\tilde{U}$ and, for ease of notation, assume that no interchange of rows is needed in the LU algorithm. Then we have factored U as

$$\begin{bmatrix} e_1 & \tilde{L} \end{bmatrix} \begin{bmatrix} e_1 & 0^T \\ & \tilde{U} \end{bmatrix}.$$

From this equation we can see the two reasons why the algorithm performs well: we have effectively decoupled the top component of y from the others and can set it to zero without introducing error. Further, the projection of b onto the range of $D-P$ is done using the diagonally dominant bidiagonal matrix $\tilde{L}$, which is guaranteed to be full rank.

In the following code fragment, we factor the matrix $\bar{U} = \tilde{L}\tilde{U}$, assuming that $\bar{U}$ is stored in the array P. We store $\tilde{U}$ in the upper triangle of P, rows 1 through n, and we store the multipliers (off-diagonal elements of the L factor) in the zeroth row of P. None of this disturbs the lower triangular factor L stored below the diagonal of P.

```
Initialize two row vectors of length  n+1:
  all entries in  ipos  are zero,
  and the i-th entry of  ind  is  i.

for i=1, ... , n,
  if |u(0,i)|  > 1,
    Interchange ind(0) with ind(i)
       and P(0,i:n) with P(i,i:n).
    Set ipos(i-1)=i.
  end if
  The pivot element is P(0,i)=-P(0,i)/P(i,i).
  Update row 0 as
    P(0,i+1:n)=P(0,i+1:n)+ P(0,i) *P(i,i+1:n).
end
```

This takes $n^2 + O(n)$ operations. The vector `ipos` is redundant but included for clarity.

We use these factors as follows to solve the linear system $ULz = b$. First we solve $\tilde{L}q = b$ in $O(n)$ operations, by using the multipliers and the permutation information:

```
Let  q  be the vector  b  reordered
   as indicated by  ind.

for i=1, ... , n,
   Set ispot=ipos(i).
   Let q(ispot)=q(ispot)+P(0,i)*q(i).
end
```

Then we solve $\tilde{U}\bar{y} = q$ using back substitution. This takes $n^2 + O(n)$ operations.

Finally, we solve $Lz = \bar{y}$, setting $z_1 = 0$. This takes $n^2 + O(n)$ operations.

Applying this algorithm to the examples in section 2 yields the results labeled $r_{improved}$ in Table 2.1. The improved algorithm yields a small residual for all of the examples. Since the residual norm divided by the condition number of $D - P$ is close to machine precision, we see that we have achieved attainable accuracy using this algorithm.

4. An application. This work was motivated by difficulties encountered in computing solutions to the telecommunications model described in Krishnan and Huebner [5]. In their model, there are n channels that serve C classes of calls. Class j calls arrive according to a Poisson process at rate λ_j, have exponential service times with mean $1/\mu_j$, and each call requires r_j channels. The classes represent different types of applications, such as voice, data, and video. The problem is to construct an admission rule that optimizes a given performance criterion, e.g., minimize the loss rate of calls. To illustrate the nature of the problem, suppose $r_1 > r_2$, and exactly r_1 channels are free when a class 1 job arrives. Accepting this job may preclude accepting several class 2 calls that might arrive soon. The class 1 job should be accepted when μ_1 is sufficiently large and λ_2 is sufficiently small so that the expected number of lost calls is less than one. This expected value depends on which calls are currently in progress (because some of them may finish soon enough to allow some class 2 calls to be admitted in the near future) as well as on which type of call is under consideration.

Krishnan and Huebner formulate this problem as a Markov decision process. This involves constructing a Markov chain to model the number of occupied channels at any time, so there is an underlying continuous-time Markov chain with states $\{0, 1, \ldots, n\}$. The nonzero elements in the rate matrix P for this chain are defined by

$$\begin{aligned} p_{i,i+r_k} &= \lambda_k && \text{for } 0 \le i \le C - r_k, && k = 1, \ldots, C, \\ p_{i,i-r_k} &= \mu_k E(m_k|i) = \tfrac{\lambda_k q(i-r_k)}{q(i)} && \text{for } r_k \le i \le n, && k = 1, \ldots, C. \end{aligned}$$

The state probabilities $q(i)$ are computed recursively using a method of Kaufman [4].

The examples of section 2 are special cases of this model with $C = 1$ class.

Let c_j be the "cost" per unit time of being in state j; e.g., the loss rate if the objective is to minimize the loss rate of calls. This model is described in continuous time, but it can be converted into a discrete-time model where transitions occur at times $1, 2, \ldots$ by "uniformizing" the model; see, e.g., Heyman and Sobel [3, section 8-7] for details. Let P be the transition matrix of the discrete-time Markov chain; P inherits the state space $\{0, 1, 2, \ldots, n\}$ and has elements p_{ij}. If some $r_j = 1$, then P is irreducible and aperiodic and has no transient states. Otherwise, some states may not be reachable (e.g., state 1 when starting empty); these states should be eliminated.

Krishnan and Huebner show that when i channels are occupied and a class j call arrives, that call should be admitted if and only if

$$t(i) < t(i + r_j),$$

where

$$t(k) = \sum_{j=0}^{n} A^{\#}_{kj} c_j, \quad k = 0, 1, \ldots, n\,.$$

From this equation (a variant of the one used by Krishnan and Huebner) we see that we need to compute the jth column of $A^{\#}$ when $c_j \neq 0$.

Example. Suppose we have $n = 100$ trunk lines, with $C = 3$ classes of traffic defined by mean arrival times (λ_i), mean holding time ($1/\mu_i$), and r_i trunks required per call as follows:

i	λ_i	μ_i	r_i
1	20	1	1
2	20	1/2	2
3	5	1/3	3

.

Using the standard algorithm, we obtain a residual of size 9.4e+12 for $j = n$. The condition number of U is computationally infinite, even though the condition number of $D - P$ is only 83. Using the algorithm from section 3, however, we obtain a residual of size 9.6e-14.

5. Conclusions. We have presented an improvement to algorithms that use the UL factors to compute quantities related to Markov chains. For a dense matrix, it requires only $O(n^2)$ additional operations compared to the standard $O(n^3)$ algorithm but improves the accuracy obtained in the results. The same approach could be used on sparse matrices arising from Markov chains.

REFERENCES

[1] W. K. Grassmann, M. I. Taksar, and D. P. Heyman, *Regenerative analysis and steady state distributions*, Oper. Res., 33 (1985), pp. 1107–1116.

[2] D. P. Heyman and D. P. O'Leary, *What is fundamental for Markov chains: First passage times, fundamental matrices, and group generalized inverses*, in Proc. Second International Workshop on Markov Chains, W. Stewart, ed., Kluwer Academic Publishers, Norwell, MA, 1995, pp. 151–161.

[3] D. P. Heyman and M. J. Sobel, *Stochastic Models in Operations Research, Vol.* I, McGraw–Hill, New York, 1982.

[4] J. S. Kaufman, *Blocking in a shared resource environment*, IEEE Trans. Comm., COM-29 (1981), pp. 1474–1481.

[5] K. R. Krishnan and F. Huebner, *Admission control for multirate CBR traffic: A Markov decision criterion*, in Teletraffic Contributions for the Information Age, V. Ramaswami and P. E. Wirth, eds., Elsevier, New York, 1997, pp. 1043–1054.

[6] C. D. Meyer, Jr., *The role of the group generalized inverse in the theory of Markov chains*, SIAM Rev., 17 (1975), pp. 443–464.

[7] C. A. O'Cinneide, *Entrywise perturbation theory and error analysis for Markov chains*, Numer. Math., 65 (1993), pp. 109–120.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 541–550, April 1998

016

QUADRATIC RESIDUAL BOUNDS FOR THE HERMITIAN EIGENVALUE PROBLEM*

ROY MATHIAS†

Abstract. Let

$$A = \begin{bmatrix} M & R \\ R^* & N \end{bmatrix} \text{ and } \tilde{A} = \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix}$$

be Hermitian matrices. Stronger and more general $O(\|R\|^2)$ bounds relating the eigenvalues of A and $\tilde{A}$ are proved using a Schur complement technique. These results extend to singular values, to eigenvalues of non-Hermitian matrices, and to generalized eigenvalues.

Key words. residual bound, Hermitian eigenvalue, Schur complement

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

PII. S0895479896310536

Let

$$A = \begin{bmatrix} M & R \\ R^* & N \end{bmatrix} \text{ and } \tilde{A} = \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} \tag{1}$$

be Hermitian matrices. Since $||A - \tilde{A}|| = ||R||$ one can bound the difference between their eigenvalues in terms of $||R||$. It is part of the folklore of numerical linear algebra that if the spectra of M and N are well separated, then a residual of size $||R||$ produces a perturbation of size $O(||R||^2)$ in the eigenvalues. The quadratic bounds that we prove are stronger, simpler, and more general than those in the literature.

For an $n \times n$ Hermitian matrix X let $\lambda_1(X) \geq \lambda_2(X) \geq \cdots \geq \lambda_n(X)$ denote its ordered eigenvalues. Throughout we shall assume that A and $\tilde{A}$ are as in (1) and that M is $m \times m$ and N is $n \times n$. Let

$$\alpha_k = \lambda_k(A) \text{ and let } \tilde{\alpha_k} = \lambda_k(\tilde{A}).$$

The eigenvalues of $\tilde{A}$ are those of M and those of N. Fix indices $i_1 < i_2 < \cdots < i_m$ such that

$$\lambda_{i_j}(\tilde{A}) = \tilde{\alpha}_{i_j} = \lambda_j(M), \qquad j = 1, 2, \ldots, m.$$

If M and N have common eigenvalues, there will be some freedom in the choice of indices i_j. We let $\sigma(X)$ denote the set of eigenvalues of X and use $\|\cdot\|$ to denote the spectral norm (often called the 2-norm by numerical analysts).

Our results are based on the following simple observation: for any $\lambda \notin \sigma(N)$

$$A - \lambda I = \begin{bmatrix} M - \lambda I & R \\ R^* & N - \lambda I \end{bmatrix}$$

is congruent to

$$\begin{bmatrix} M - \lambda I - R(N - \lambda I)^{-1} R^* & 0 \\ 0 & N - \lambda I \end{bmatrix} = \tilde{A} - \lambda I - E(\lambda),$$

*Received by the editors October 11, 1996; accepted for publication (in revised form) by R. Bhatia April 11, 1997.

http://www.siam.org/journals/simax/19-2/31053.html

†College of William and Mary, Department of Mathematics, Williamsburg, VA 23185 (mathias@math.wm.edu). This research was supported in part by NSF grant DMS-9504795.

where

$$E(\lambda) \equiv \begin{bmatrix} R(N-\lambda I)^{-1}R^* & 0 \\ 0 & 0 \end{bmatrix},$$

and so $A - \lambda I$ and $\tilde{A} - \lambda I - E(\lambda)$ have the same inertia.

LEMMA 1. *If $\alpha_k \notin \sigma(N)$, then*

$$\alpha_k = \lambda_k(A) = \lambda_k(\tilde{A} - E(\alpha_k)) \tag{2}$$

while if $\tilde{\alpha}_k \notin \sigma(N)$, then

$$\tilde{\alpha}_k = \lambda_k(\tilde{A}) = \lambda_k(A + E(\tilde{\alpha}_k)). \tag{3}$$

Proof. Since $\lambda_k(A - \alpha_k I) = 0$ and because $A - \alpha_k I$ and $\tilde{A} - \alpha_k I - E(\alpha_k)$ have the same inertia,

$$0 = \lambda_k(\tilde{A} - \alpha_k I - E(\alpha_k)) = \lambda_k(\tilde{A} - E(\alpha_k)) - \alpha_k,$$

which implies (2). For (3) one can show that $A + E(\tilde{\alpha}_k) - \tilde{\alpha}_k I$ and $\tilde{A} - \tilde{\alpha}_k I$ are congruent, and so

$$0 = \lambda_k(\tilde{A} - \tilde{\alpha}_k I) = \lambda_k(A + E(\tilde{\alpha}_k) - \tilde{\alpha}_k I) = \lambda_k(A + E(\tilde{\alpha}_k)) - \tilde{\alpha}_k$$

as required. ∎

Since $\lambda_k(A) = \lambda_k(\tilde{A} - E(\alpha_k))$ we can bound $|\alpha_k - \tilde{\alpha}_k|$ by bounding $||E(\alpha_k)||$. This will give us the desired $O(||R||^2)$ bound. The rest of the paper is devoted to quantifying this observation.

We define a number of measures of separation from the spectrum of N:

$$\begin{aligned} \delta_k &\equiv \min_{i=1,2,\ldots,n} |\alpha_k - \lambda_i(N)|, \\ \tilde{\delta}_k &\equiv \min_{i=1,2,\ldots,n} |\tilde{\alpha}_k - \lambda_i(N)|, \\ \delta &\equiv \min_{j=1,2,\ldots,m} \delta_j = \min_{j=1,\ldots,m \;\; k=1,\ldots,n} |\lambda_{i_j}(A) - \lambda_k(N)|, \\ \text{and} \qquad \tilde{\delta} &\equiv \min_{j=1,2,\ldots,m} \tilde{\delta}_j = \min_{j=1,\ldots,m \;\; i=1,\ldots,n} |\lambda_j(M) - \lambda_i(N)|. \end{aligned}$$

We could use just δ and $\tilde{\delta}$ and the resulting bounds would be weaker, though slightly simpler.

THEOREM 1. *If $\alpha_k \notin \sigma(N)$, then*

$$|\alpha_k - \tilde{\alpha}_k| \leq \delta_k^{-1} ||R||^2 \tag{4}$$

while if $\tilde{\alpha}_k \notin \sigma(N)$, then

$$|\alpha_k - \tilde{\alpha}_k| \leq \tilde{\delta}_k^{-1} ||R||^2. \tag{5}$$

Proof. From (2) and the definition of δ_k

$$\begin{aligned} |\alpha_k - \tilde{\alpha}_k| &= |\lambda_k(A) - \lambda_k(\tilde{A})| \\ &= |\lambda_k(\tilde{A} - E(\alpha_k)) - \lambda_k(\tilde{A})| \\ &\leq ||E(\alpha_k)|| \\ &= ||R(N - \alpha_k I)^{-1} R^*|| \\ &\leq \delta_k^{-1} ||R||^2 \end{aligned}$$

which is (4). The inequality (5) can be proved in the same way. ∎

The bound (5) gives us a different bound on $|\alpha_k - \tilde{\alpha}_k|$ for each eigenvalue $\tilde{\alpha}_k$ of M. So, for example, if $\tilde{\alpha}_k$ is well separated from the spectrum of N, then $|\alpha_k - \tilde{\alpha}_k|$ is $O(||R||^2)$ even though M and N may have common eigenvalues.

We can maximize the right-hand sides of (4) and (5) over k and obtain weaker but more familiar bounds.

COROLLARY 1. *If* $||R|| < \delta$, *then*

$$\max_{k=1,2,\ldots,m} |\alpha_{i_k} - \lambda_k(M)| \leq \delta^{-1}||R||^2. \tag{6}$$

Even without the condition $||R|| < \delta$,

$$\max_{k=1,2,\ldots,m+n} |\alpha_k - \tilde{\alpha}_k| \leq \tilde{\delta}^{-1}||R||^2. \tag{7}$$

The inequality (6) gives an affirmative answer to the question asked by Sun in [12, section 5.2].[1]

These results are an improvement over the results given by Stewart [10, Theorem 3.12] (or [11]) and Sun [12, Corollary 3.4] in a number of ways. First, their results require that m eigenvalues of A lie outside $[\lambda_n(N) - \delta, \lambda_1(N) + \delta]$. That is, m eigenvalues of A must be well separated from the convex hull of $\sigma(N)$, while here we merely require that they be well separated from $\sigma(N)$. Second, their bounds contain a factor $(1 - \rho^2)^{-1}$, where $\rho = ||R||\delta$. This factor is always greater than one, though in most applications it would be close to one. Our bound contains no such factor.[2] Third, our proof is simpler—the proofs due to Stewart and Sun use theorems on the perturbation of invariant subspaces and a perturbation result for matrices that are similar to Hermitian matrices. Fourth, as observed above, our results are valid even when M and N have common eigenvalues. Finally, our results are directly applicable in a situation that often arises in numerical linear algebra—when one knows or has bounds on the spectra of M and N but not that of A. Here our bounds in terms of $\tilde{\delta}_k$ and $\tilde{\delta}$ are directly applicable and will give stronger bounds than using the information about $\sigma(M)$ and $\sigma(N)$ to deduce information about $\sigma(A)$ and $\sigma(N)$ and then use this to obtain $0(||R||^2)$ bounds involving δ.

The idea of shifting A by an eigenvalue, applying a congruence, and then looking at the resulting Schur complement is not new. It has been used a number of times by Parlett [9, sections 9.5.1, 10.1, 10.4, 10.5]. It seems that our results are different from the results in [9] where the primary focus is on multiple eigenvalues. Schur complements arise in the work of Mathias and Stewart in analyzing the perturbation of eigenvalues of graded matrices.

The key idea in this paper, which is summarized in (2)–(3), is that if one has a special perturbation of the form (1), then the resulting perturbation in the eigenvalues of A is the same as that if one makes a much smaller perturbation. That is, because of the special structure of the perturbation one can use *standard perturbation theory* to derive stronger perturbation bounds. This idea is not new: Stewart [11] has used it, and it is an underlying idea in the work of Eisenstat and Ipsen [3, 4], although it is not identified by them.

The relation (3) gives a perturbation of size $O(\|R\|^2)$ of A such that the kth eigenvalue of $\tilde{A}$ is the the kth eigenvalue of the slightly perturbed A. The perturbation

[1] Bhatia [2] and Mathias [7] have answered the question in [12, section 5.1] in the negative.

[2] Actually the factor $(1 - \rho^2)^{-1}$ in [10] can be reduced to $(1 + \rho^2)^{1/2}$, which does not blow up as $p \to 1$, by noting that $||P(I + P^*P)^{1/2}|| \leq \rho$ implies $||P|| \leq \rho$ because $||(I + P^*P)^{-1/2}|| \leq 1$. At any rate, our result is cleaner for not having the factor $(1 + \rho^2)^{1/2}$.

is different for each index. The inequality below, which is from the proof of Theorem 3, may be viewed as an extension of (3) giving two perturbations that yield $O(\|R\|^2)$ bounds on all the eigenvalues.

$$\lambda_i(A - \tilde{\delta}^{-1}W) \le \lambda_i(\tilde{A}) \le \lambda_i(A + \tilde{\delta}^{-1}W), \quad i = 1, 2, \ldots, m+n,$$

where $W = RR^* \oplus R^*R$. This inequality suggests that one can extend the results in this section to a wider class of norms, and we do this in the next section, although the bounds (4)–(7) should be sufficient for most purposes.

We conclude the paper by showing how this Schur complement technique can be applied to the non-Hermitian eigenvalue problem and the singular value problem. One can also apply it to the generalized eigenvalue problem, but it is necessary to introduce a fair amount of new notation to state the strongest possible result so we do not apply the Schur complement technique to this situation.

Extension to unitarily invariant norms. Let Φ denote a symmetric gauge function and let $||\cdot||_\Phi$ denote the corresponding unitarily invariant norm. It is known that every unitarily invariant norm is equal to $||\cdot||_\Phi$ for some Φ. (For a proof of this fact see, for example, [10, Theorem II.3.6] or [5, Chapter 5].) Also, if $X = X^*$ is $k \times k$, then

$$||X||_\Phi = \Phi(\lambda_1(X), \ldots, \lambda_k(X)).$$

Consider a symmetric gauge function Φ on $\mathbb{R}^k$. It induces, in a natural way, a symmetric gauge function Φ_r on $\mathbb{R}^r$ for any positive integer r. If $r \le k$, then

$$\Phi_r(x) = \Phi(x_1, \ldots, x_r, 0, \ldots, 0),$$

while if $r \ge k$, then

$$\Phi_r(x) = \max_{1 \le i_1 < i_2 < \cdots < i_k \le r} \Phi(x_{i_1}, \ldots, x_{i_k}).$$

In this way we may think of Φ acting on $\mathbb{R}^r$ for any r. This simplifies the statements and proofs of some of the results in this section.

We shall make frequent use of *Weyl's monotonicity principle* (see, e.g., [10, Corollary IV.4.9, and the subsequent discussion] or [5, Corollary 4.3.3]): if $X \le Y$ are $k \times k$, then

$$\lambda_i(X) \le \lambda_i(Y), \quad i = 1, 2, \ldots, k.$$

We will also need the Lidskii–Wielandt bound which is the next result. See [10, Theorem IV.4.8] or [1] for the traditional proof via Wielandt's min-max theorem or see [6] for a much more elementary proof using Weyl's monotonicity principle.

THEOREM 2. *Let X and Y be $n \times n$ Hermitian matrices, and let Φ be a symmetric gauge function on $\mathbb{R}^n$; then*

$$\Phi(\lambda_1(X) - \lambda_1(Y), \ldots, \lambda_n(X) - \lambda_n(Y)) \le ||X - Y||_\Phi.$$

We will not use Theorem 2 directly; rather we will use Lemma 2 which is easily deduced from it.

LEMMA 2. *Let B and C be $n \times n$ Hermitian matrices and let $\alpha_1, \alpha_2, \ldots, \alpha_n$ and $\tilde{\alpha_1}, \tilde{\alpha_2}, \ldots, \tilde{\alpha_n}$ be real numbers. Let Φ be any symmetric gauge function on $\mathbb{R}^n$. If for some sequence $i_1 < i_2 < \cdots < i_k$ we have*

$$\lambda_{i_j}(B) \le \alpha_{i_j} \le \lambda_{i_j}(C) \text{ and } \lambda_{i_j}(B) \le \tilde{\alpha}_{i_j} \le \lambda_{i_j}(C), \quad j = 1, 2, \ldots, k,$$

then

$$\Phi(\alpha_{i_1} - \tilde{\alpha}_{i_1}, \ldots, \alpha_{i_k} - \tilde{\alpha}_{i_k}, 0, \ldots, 0) \leq ||B - C||_\Phi.$$

Proof. Let $\gamma_i = |\lambda_i(C) - \lambda_i(B)|$. The bounds on α_{i_j} and $\tilde{\alpha}_{i_j}$ imply $|\alpha_{i_j} - \tilde{\alpha}_{i_j}| \leq \gamma_i$. Thus,

$$\begin{aligned}\Phi(\alpha_{i_1} - \tilde{\alpha}_{i_1} \ldots, \alpha_{i_k} - \tilde{\alpha}_{i_k}) &\leq \Phi(\gamma_{i_1}, \ldots, \gamma_{i_k}) \\ &\leq \Phi(\gamma_1, \ldots, \gamma_n) \\ &= \Phi(\lambda_1(B) - \lambda_1(C), \ldots, \lambda_n(B) - \lambda_n(C)) \\ &= ||B - C||_\Phi. \qquad \square\end{aligned}$$

We can now generalize Corollary 1 to all unitary invariant norms.

THEOREM 3. *For any symmetric gauge function* Φ

$$\Phi(\alpha_{i_1} - \tilde{\alpha}_{i_1}, \ldots, \alpha_{i_m} - \tilde{\alpha}_{i_m}) \leq 2\delta^{-1}||RR^*||_\Phi, \tag{8}$$

$$\Phi(\alpha_1 - \tilde{\alpha}_1, \ldots, \alpha_{i_m} - \tilde{\alpha}_{i_m}) \leq 2\tilde{\delta}^{-1}||RR^*||_\Phi, \tag{9}$$

and

$$\Phi(\alpha_1 - \tilde{\alpha}_1, \ldots, \alpha_{m+n} - \tilde{\alpha}_{m+n}) \leq 2\tilde{\delta}^{-1}||RR^* \oplus R^*R||_\Phi. \tag{10}$$

These results can easily be stated in the notation used in [12]. For example, (10) would be

$$\|\mathrm{diag}(\alpha_1 - \tilde{\alpha}_1, \ldots, \alpha_{m+n} - \tilde{\alpha}_{m+n})\|_\Phi \leq 2\tilde{\delta}^{-1}||RR^* \oplus R^*R||_\Phi.$$

Proof. We will prove (9) and (10) only; (8) can be proved in exactly the same way as (9). Since

$$-\delta^{-1}(RR^* \oplus 0) \leq E(\alpha_k) \leq \delta^{-1}(RR^* \oplus 0),$$

the basic relation (3) gives us

$$\tilde{\alpha}_{i_k} = \lambda_{i_k}(A + E(\alpha_k)) \leq \lambda_{i_k}(A + \delta^{-1}RR^* \oplus 0)$$

and

$$\tilde{\alpha}_{i_k} = \lambda_{i_k}(A + E(\alpha_k)) \geq \lambda_{i_k}(A - \delta^{-1}RR^* \oplus 0),$$

which together imply

$$\lambda_{i_k}(A - \delta^{-1}RR^* \oplus 0) \leq \tilde{\alpha}_{i_k} \leq \lambda_{i_k}(A + \delta^{-1}RR^* \oplus 0). \tag{11}$$

Since $RR^* \geq 0$, Weyl's monotonicity principle immediately yields exactly the same bound on α_{i_k}:

$$\lambda_{i_k}(A - \delta^{-1}RR^* \oplus 0) \leq \alpha_{i_k} \leq \lambda_{i_k}(A + \delta^{-1}RR^* \oplus 0). \tag{12}$$

The bounds (11) and (12) together with Lemma 2 imply (9). A similar proof shows that for $i \notin \{i_1, \ldots, i_m\}$

$$\lambda_i(A - \tilde{\delta}^{-1}(0 \oplus R^*R)) \leq \tilde{\alpha}_i \leq \lambda_i(A + \tilde{\delta}^{-1}(0 \oplus RR^*)).$$

Thus, for any $i \in \{1, 2, \ldots, m+n\}$

$$\lambda_i(A - \tilde{\delta}^{-1}(RR^* \oplus R^*R)) \leq \tilde{\alpha}_i \leq \lambda_i(A + \tilde{\delta}^{-1}(RR^* \oplus R^*R))$$

and

$$\lambda_i(A - \tilde{\delta}^{-1}(RR^* \oplus R^*R)) \leq \alpha_i \leq \lambda_i(A + \tilde{\delta}^{-1}(RR^* \oplus R^*R)).$$

Again, Lemma 2 implies (10). □

One would like to remove the factor 2 from these bounds, but we have not been able to. It is easy to show that if we fix M and N and let $\|R\| \to 0$, then the bounds are valid asymptotically without the factor 2. An examination of the proof of the theorem shows that we can replace $2\|RR^*\|_\Phi$ in (8) and (9) by

$$\Phi(\lambda_1(RR^*), \lambda_1(RR^*), \lambda_2(RR^*), \lambda_2(RR^*), \ldots),$$

where Φ has m arguments. This is slightly stronger but is still not the ideal result. One can completely remove the factor 2 from the bounds in (8)–(10) by imposing a more stringent condition on the spectra of $\tilde{A}$ and A. Notice that the condition for the validity of (14) is not symmetric in M and N.

THEOREM 4. *Let Φ be a symmetric gauge function. If exactly n of the eigenvalues of A lie outside $(\lambda_m(M) - \delta, \lambda_1(M) + \delta)$, then*

$$\Phi(\alpha_i - \tilde{\alpha}_{i_1}, \ldots, \alpha_{i_m} - \tilde{\alpha}_{i_m}) \leq \delta^{-1}||RR^*||_\Phi. \tag{13}$$

If all the eigenvalues of N lie outside $(\lambda_m(M) - \tilde{\delta}, \lambda_1(M) + \tilde{\delta})$, then

$$\Phi(\alpha_{i_1} - \tilde{\alpha}_{i_1}, \ldots, \alpha_{i_m} - \tilde{\alpha}_{i_m}) \leq \tilde{\delta}^{-1}||RR^*||_\Phi. \tag{14}$$

If $\lambda_1(N) = \lambda_m(M) - \tilde{\delta}$, then

$$\begin{aligned} \alpha_i &\geq \tilde{\alpha}_i, \qquad i = 1, \ldots, m, \\ \alpha_i &\leq \tilde{\alpha}_i, \qquad i = m+1, \ldots, m+n, \end{aligned}$$

and

$$\Phi(\alpha_1 - \tilde{\alpha}_1, \ldots, \alpha_{m+n} - \tilde{\alpha}_{m+n}) \leq ||RR^* \oplus R^*R||_\Phi. \tag{15}$$

Proof. We shall prove only (14); the rest of the theorem can be proved in the same way. We may apply a unitary similarity of the form

$$\begin{bmatrix} I & 0 \\ 0 & U \end{bmatrix}$$

to A and $\tilde{A}$ without changing any of the quantities in the theorem. Thus, without loss of generality we may assume that

$$A = \begin{bmatrix} M & R_1 & R_2 \\ R_1^* & N_1 & 0 \\ R_2^* & 0 & N_2 \end{bmatrix},$$

where the spectrum of N_1 is in $(-\infty, \lambda_m(M) - \tilde{\delta}]$ and that of N_2 is in $[\lambda_1(M) + \tilde{\delta}, \infty)$. Thus, for any $\alpha \in [\lambda_m(M), \lambda_1(M)]$

$$0 \leq R_2(N_2 - \alpha I)^{-1}R_2^* \leq \tilde{\delta}^{-1}R_2R_2^*$$

and

$$-\tilde{\delta}^{-1}R_1R_1^* \le R_1(N_1-\alpha I)^{-1}R_1^* \le 0.$$

Since

$$R(N-\alpha I)^{-1}R^* = R_1(N_1-\alpha I)^{-1}R_1^* + R_2(N_2-\alpha I)^{-1}R_2^*,$$

if $\alpha \in [\lambda_m(M), \lambda_1(M)]$, it follows that

$$-\delta^{-1}R_1R_1^* \oplus 0 \le E(\alpha) \le \delta^{-1}R_2R_2^* \oplus 0. \tag{16}$$

Take any α_{i_k}. Then from (2), Weyl's monotonicity principle, and (16) we have

$$\lambda_{i_k}(\tilde{A}-\tilde{\delta}^{-1}R_2R_2^*\oplus 0) \le \alpha_{i_k} \le \lambda_{i_k}(\tilde{A}+\tilde{\delta}^{-1}R_1R_1^*\oplus 0)$$

and, of course,

$$\lambda_{i_k}(\tilde{A}-\tilde{\delta}^{-1}R_2R_2^*\oplus 0) \le \tilde{\alpha}_{i_k} \le \lambda_{i_k}(\tilde{A}+\tilde{\delta}^{-1}R_1R_1^*\oplus 0).$$

Lemma 2 now implies

$$\begin{aligned}\Phi(\alpha_{i_1}-\tilde{\alpha}_{i_1},\ldots,\alpha_{i_m}-\tilde{\alpha}_{i_m}) &\le ||(\tilde{A}-\tilde{\delta}^{-1}R_1R_1^*\oplus 0)-(\tilde{A}-\tilde{\delta}^{-1}R_2R_2^*\oplus 0)||_\Phi \\ &= ||\tilde{\delta}^{-1}(R_1R_1^*+R_2R_2^*)\oplus 0||_\Phi \\ &= \tilde{\delta}^{-1}||RR^*||_\Phi\end{aligned}$$

as desired. □

Notice that our bounds are in terms of $||RR^*||_\Phi$ rather than $||R||_\Phi^2$. This is an advantage, since one can show that if Φ is normalized, that is, if

$$\Phi(1,0,0,\ldots,0)=1,$$

as is the case for the spectral, Frobenius, trace, Schatten-p and Ky–Fan-k norms, then

$$||RR^*||_\Phi \le \|R\|\,\|R\|_\Phi \le ||R||_\Phi^2.$$

The first inequality shows that our bound is stronger than Sun's [12, Corollary 3.4] which has $\|R\|\|R\|_\Phi$ in the bound, as well as an additional factor $(1-\rho^2)^{-1/2}$.

Extension to non-Hermitian matrices. One can apply these techniques to non-Hermitian matrices. However, the results are rather clumsy as shown below.

THEOREM 5. *Let*

$$A = \begin{bmatrix} M & R \\ S & N \end{bmatrix}$$

be a diagonalizable $(m+n)\times(m+n)$ matrix. Let $\tilde{\lambda} \notin \sigma(N)$ be an eigenvalue of M and let X be a matrix that diagonalizes A. Then there is an eigenvalue λ of A such that

$$|\lambda-\tilde{\lambda}| \le \kappa(X)||R||\;||S||\;||(N-\tilde{\lambda}I)^{-1}||. \tag{17}$$

Proof. The singular matrix

$$A + \begin{bmatrix} R(N-\tilde{\lambda}I)^{-1}S & 0 \\ 0 & 0 \end{bmatrix} - \tilde{\lambda}I$$

is equivalent to

$$\begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix} - \tilde{\lambda} I.$$

Consequently, $\tilde{\lambda}$ is an eigenvalue of

$$A + [R(N - \tilde{\lambda} I)^{-1} S] \oplus 0.$$

The result now follows from this and standard results on the perturbation of diagonalizable matrices, e.g., [10, Theorem IV.3.3]. □

The bound (17) may be useful when A is normal and $N = \nu$ is 1×1. In this case X can be chosen unitary, with $\kappa(X) = 1$, and also $||R|| = ||S||$ and $(N - \tilde{\lambda} I)^{-1} = (\nu - \tilde{\lambda})^{-1}$, so we have

$$|\lambda - \tilde{\lambda}| \leq \frac{||R||^2}{|\nu - \tilde{\lambda}|}$$

which is analogous to the Hermitian case. Of course ν is also an estimate of an eigenvalue of A; there is an eigenvalue λ_ν of A such that

$$|\nu - \lambda_\nu| \leq ||R||^2 ||(M - \nu I)^{-1}||.$$

This is a quadratic version of the residual bound [10, Theorem IV.3.2].

Application to residual bounds for singular values. The results for eigenvalues imply similar results for singular values. For example, let

$$A = \begin{bmatrix} M & R \\ S & N \end{bmatrix}$$

be a general matrix—we do not require M and N to be square, and let

$$\tilde{A} = \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix}.$$

Since the Jordan–Wielandt matrix

$$\begin{bmatrix} 0 & 0 & M & R \\ 0 & 0 & S & N \\ M^* & S^* & 0 & 0 \\ R^* & N^* & 0 & 0 \end{bmatrix}$$

is permutation similar to

$$\begin{bmatrix} 0 & M & 0 & R \\ M^* & 0 & S^* & 0 \\ 0 & S & 0 & N \\ R^* & 0 & N^* & 0 \end{bmatrix},$$

the results in the first two sections imply a number of bounds on the difference between the singular values of A and $\tilde{A}$. Theorem 1, for example, yields

$$\max |\sigma_i(A) - \sigma_i(\tilde{A})| \leq \frac{\max\{||R||^2, ||S||^2\}}{\min_{j,k} |\sigma_j^2(M) - \sigma_k^2(N)|}.$$

One may also be interested in the relative error in approximating the singular values of

$$A = \begin{bmatrix} M & R \\ 0 & N \end{bmatrix}$$

by those of

$$\tilde{A} = \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix},$$

where in this case M and N are again square. We consider the case $\sigma_{\min}(M) > \sigma_1(N)$. (We could omit this restriction at the cost of a considerably more complicated bound.) Then

$$AA^* = \begin{bmatrix} M & R \\ 0 & N \end{bmatrix} \begin{bmatrix} M^* & 0 \\ R^* & N^* \end{bmatrix} = \begin{bmatrix} MM^* + RR^* & RN^* \\ NR^* & NN^* \end{bmatrix}.$$

Let σ_{i+m} be the $i+m$th singular value of A. Then $\lambda_{m+i}(AA^* - \sigma_{i+m}^2 I) = 0$, and as before this implies that

$$\lambda_{m+i} \begin{bmatrix} MM^* + RR^* - \sigma_{i+m} I & 0 \\ 0 & NN^* - \sigma_{i+m} I - NR^*(MM^* + RR^* - \sigma_{i+m}^2 I)^{-1} RN^* \end{bmatrix} = 0.$$

Because $\sigma_m(M) > \sigma_1(N)$ it follows that $MM^* + RR^* - \sigma_{i+m}^2 I$ is positive definite.[3] So

$$\lambda_i(NN^* - NR^*(MM^* + RR^* - \sigma^2 I)^{-1} RN^* - \sigma_{i+m}^2 I) = 0$$

and hence

$$\lambda_i(N[I - R^*(MM^* + RR^* - \sigma_{i+m}^2 I)^{-1} R]N^*) = \sigma_{m_i}^2.$$

Since

$$0 \le R^*(MM^* + RR^* - \sigma^2 I)^{-1} R \le (\sigma_{\min}^2 - \sigma_{i+m}^2)^{-1} ||R||^2,$$

it follows that

$$1 \ge \frac{\sigma_{i+m}^2}{\lambda_i(NN^*)} \ge \left(1 - \frac{||R||^2}{\sigma_{\min}^2(M) - \sigma_{i+m}^2}\right).$$

Of course $\lambda_i(NN^*) = \sigma_i^2(N)$.

To summarize, we have shown

$$1 \ge \frac{\sigma_{m+i}^2}{\sigma_i(N)} \ge \left(1 - \frac{||R||^2}{\sigma_{\min}^2(M) - \sigma_i^2}\right)^{1/2}.$$

[3] If $MM^* + RR^* - \sigma_{m+i}^2 I$ were not positive definite, then $\sigma_{min}^2(M) \le \sigma_{m+i}^2$ and so $NN^* - \sigma_{m+i}^2 I < 0$. That is, an $n \times n$ principal submatrix of $AA^* - \sigma_{m+i}^2 I$ is negative definite and so $AA^* - \sigma_{m+i}^2 I$ must have at least n negative values. This contradicts the fact that since $\lambda_{m+i}(AA^*) = \sigma_{m+i}^2$ the matrix $AA^* - \sigma_{m+i}^2 I$ has exactly $(m+n) - (m+i) = (n-i)$ negative eigenvalues.

This is always at least slightly stronger than [8, equation (3.3)], and is considerably stronger when $\sigma_{\min}(M)/\sigma_1(N)$ is close to one, but $\sigma_{\min}(M)/\sigma_i(N)$ is far from one. The result in [8] was proved in a very different way.

Using similar techniques one can show

$$\left(1 + \frac{||R||^2 \cdot \rho^2}{\sigma_i^2(A) - \sigma_1^2(N)}\right)^{1/2} \geq \frac{\sigma_i(A)}{\sigma_i([M\ R])} \geq 1.$$

This bound is also stronger than [8, equation (3.4)]. Not only is our "gap" larger, but there is also the additional factor ρ^2, where

$$\rho = \frac{||E||}{\sigma_{\min}(M)} < 1,$$

multiplying the $||R||^2$ term.

REFERENCES

[1] R. Bhatia, *Perturbation Bounds for Matrix Eigenvalues*, Pitman Res. Notes Math. Ser. 162, Longman Scientific and Technical, New York, 1987.

[2] R. Bhatia, *On residual bounds for eigenvalues*, Indian J. Pure Appl. Math., 23 (1992), pp. 865–866.

[3] S. C. Eisenstat and I. C. F. Ipsen, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.

[4] S. C. Eisenstat and I. C. F. Ipsen, *Relative Perturbation Bounds for Eigenvalues and Eigenvectors of Diagonalizable Matrices*, Tech. report 96-6, Department of Mathematics, North Carolina State University, Raleigh, NC, 1996.

[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 1985.

[6] C.-K. Li and R. Mathias, *The Lidskii–Mirsky–Wielandt theorem—Additive and multiplicative versions*, Numer. Math., submitted.

[7] R. Mathias, *Residual Bounds for Hermitian Matrices with Unfavorable Eigenvalue Distribution*, 1993, unpublished manuscript.

[8] R. Mathias and G. W. Stewart, *A block QR algorithm and the singular value decomposition*, Linear Algebra Appl., 182 (1993), pp. 91–100.

[9] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[10] G. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, MA, 1990.

[11] G. W. Stewart, *Two simple residual bounds for the eigenvalues of a Hermitian matrix*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 205–208.

[12] J.-G. Sun, *Eigenvalues of Rayleigh quotient matrices*, Numer. Math., 59 (1991), pp. 603–614.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 551–555, April 1998

017

THE RELATION BETWEEN THE QR AND LR ALGORITHMS*

HONGGUO XU†

Abstract. For an Hermitian matrix the QR transform is diagonally similar to two steps of the LR transforms. Even for non-Hermitian matrices the QR transform may be written in rational form.

Key words. QR algorithm, LR algorithm, triangular factorization, Cholesky factorization

AMS subject classification. 65F15

PII. S0895479896299937

1. Summary. In this section we assume that the reader is familiar with LR and QR. The transformations are presented in the next section. A slight variation of the LR algorithm that is suitable for positive definite Hermitian matrices is the Cholesky (LR) algorithm: $A = CC^*$ is mapped into $\hat{A} = C^*C$.

In the positive definite Hermitian case, two steps of Cholesky yield the same matrix as one step of QR. At first glance this is surprising (how can LR produce a unitary similarity?) and the proof is sometimes given as an exercise in textbooks; see [4], [1], and [9]. Students are often left with the (false) impression that positive definiteness is essential.

In recent years understanding of these algorithms has improved. Both the LR and QR algorithms are instances of GR algorithms [10]. For all such algorithms k steps applied to A are equivalent to a similarity driven by a factorization of A^k:

$$A^k = G_k R_k, \qquad A \to G_k^{-1} A G_k. \tag{1.1}$$

Consequently, two steps of LR on A is equivalent to a similarity driven by $A^2 : A^2 = LU, A \to L^{-1}AL$. On the other hand, one step of QR on A is equivalent (up to a diagonal similarity) to a similarity driven by A^*A:

$$A = QR, \qquad A^*A = R^*R(= LD^2L^*), \qquad A \to Q^*AQ = RAR^{-1}.$$

If A is Hermitian, then $A^*A = A^2$ and two steps of LR must be equivalent to one of QR. Despite these remarks it is still interesting to see the equivalence in detail, and that is the topic of section 2.

The catch is that LR can break down so the more careful statement is that two LR steps (if they exist) are equivalent to one QR step (which always exists). What is of more than passing interest is that LR is entirely rational in operation whereas QR requires square roots and is not rational. The remarks made above show that these square roots in QR are somehow not essential; QR may be better thought of as LR driven by A^*A. It is this viewpoint that leads to the various root-free QR algorithms that have been so successful for symmetric tridiagonal matrices. Four versions are described in [6] and an even faster one appeared recently in [3].

*Received by the editors March 1, 1996; accepted for publication (in revised form) by J. Varah April 17, 1997.

http://www.siam.org/journals/simax/19-2/29993.html

†Department of Mathematics, Fudan University, Shanghai, 200433, P. R. China. Current address: Fakultät für Mathematik, TU Chemnitz-Zwickau, D-09107 Chemnitz, Germany (hxu@mathematik.tu-chemnitz.de). This work was supported in part by the National Natural Science Foundation of China.

For non-Hermitian matrices there is a similar rational version of QR and it is described in section 3.

We follow Householder conventions for notation except for denoting the conjugate transpose of F by F^* instead of F^H. We ignore shifts because they complicate the analysis and add nothing to a theoretical paper.

2. Connection of QR to LR. Two algorithms that play an important role in matrix eigenvalue computations are the LR and QR algorithms. The former was discovered by Rutishauser in 1958 [7] and the latter developed by Francis in 1959–1960 [2]. A formal derivation of QR was given in [5]. Both algorithms have been widely studied and good references are [9], [8], and [6].

Recall the basic decompositions.

Triangular factorization (LU). If and only if $B \in \mathbf{C}^{n\times n}$ has nonzero leading principal minors of orders $1, 2, \ldots, n-1$, then B has a unique decomposition

$$B = LDU,$$

where L is unit lower triangular, D is diagonal, and U is unit upper triangular.

Gram-Schmidt factorization (QR). All $B \in \mathbf{C}^{n\times n}$ may be written

$$B = QR,$$

where $Q^* = Q^{-1}$ and R is upper triangular with nonnegative diagonal entries. The factorization is unique if and only if the columns of B are linearly independent.

From the basic factorizations come the basic transforms.

LR transform. If $B = LDU$, then its LR transform is defined by

$$\overset{\circ}{B} = DUL = L^{-1}BL = (DU)B(DU)^{-1}.$$

Here is the irritating ambiguity in LR; the definition $\overset{\circ}{B} = ULD$ would be equally legitimate. For theoretical purposes one could consider the equivalence class of all diagonal similarities on a given matrix.

QR Transform. If $B = QR$ (uniquely), then its QR transform is defined by

$$\hat{B} = RQ = Q^*BQ = RBR^{-1}.$$

Remark 1. If A is Hermitian and positive definite, then $A = LD^2L^*$ and its *Cholesky transform* is given by $A' = DL^*LD$. Note that

$$A' = D^{-1}\ \overset{\circ}{A}\ D$$

is a diagonal similarity transformation but uses square roots. LR destroys the Hermitian property but only by a diagonal similarity.

Denote the Cholesky transform of A' by A'' and the LR transform of $\overset{\circ}{A}$ by $\overset{\circ\circ}{A}$. If A is Hermitian and positive definite, then $A'' = \hat{A}$: two steps of Cholesky equal one of QR. However, the positive definite property is not essential as the following result shows.

All L's are unit lower triangular and all D's are diagonal and real. For completeness we include all the diagonal matrices.

THEOREM 2.1. *If A is Hermitian, and permits triangular factorization, then $\overset{\circ\circ}{A}$ is diagonally similar to $\hat{A}$.*

Proof. By hypothesis $A = L_1 D_1 L_1^*$ and so

$$\overset{\circ}{A} = D_1 L_1^* L_1.$$

Since $L_1^* L_1$ is positive definite it permits triangular factorization

$$L_1^* L_1 = L_2 D_2^2 L_2^* \qquad (D_2 \text{ positive}). \tag{2.1}$$

Consequently, the triangular factorization of $\overset{\circ}{A}$ is

$$\overset{\circ}{A} = (D_1 L_2 D_1^{-1})(D_1 D_2^2) L_2^*.$$

Thus,

$$\begin{aligned} \overset{\circ\circ}{A} &= D_1 D_2^2 L_2^* D_1 L_2 D_1^{-1} \\ &= (D_1 D_2) M (D_1 D_2)^{-1}, \end{aligned}$$

where

$$M := D_2 L_2^* D_1 L_2 D_2.$$

It remains to show that M is similar to $\hat{A}$ with a diagonal unitary transformation. Rewrite (2.1) as

$$I = (L_1^{-*} L_2 D_2)(D_2 L_2^* L_1^{-1}).$$

Since D_2 is real

$$Q = L_1^{-*} L_2 D_2 \tag{2.2}$$

is unitary. Use $Q = Q^{-*}$ to obtain another triangular factorization of Q

$$Q = L_1 L_2^{-*} D_2^{-1}. \tag{2.3}$$

Now use Q to rewrite M as

$$\begin{aligned} M &= (D_2 L_2^* L_1^{-1})(L_1 D_1 L_1^*)(L_1^{-*} L_2 D_2) \\ &= Q^* A Q. \end{aligned} \tag{2.4}$$

Finally, using (2.3),

$$\begin{aligned} A &= L_1 D_1 L_1^* \\ &= (L_1 L_2^{-*} D_2^{-1})(D_2 L_2^* D_1 L_1^*) \\ &= Q \operatorname{sign}(D_1) \cdot \operatorname{sign}(D_1) D_2 L_2^* D_1 L_1^* \\ &= Q \operatorname{sign}(D_1) R \end{aligned}$$

reveals the QR factorization of A since R has nonnegative diagonal. By (2.4)

$$\begin{aligned} \hat{A} &= \operatorname{sign}(D_1) M \operatorname{sign}(D_1) \\ &= \operatorname{sign}(D_1)(D_1 D_2)^{-1} \overset{\circ\circ}{A} (D_1 D_2) \operatorname{sign}(D_1)^{-1}, \end{aligned}$$

as claimed. ∎

Remark 2. When the LR transform is to be applied to an Hermitian matrix it is possible to modify the algorithm so that the Hermitian property is restored after two steps. In the notation used above

$$\overset{\circ}{A} = D_1 L_1^* L_1 = D_1 (L_2 D_2)(L_2 D_2)^*$$

and one then redefines $\overset{\circ\circ}{A}$ by

$$\overset{\circ\circ}{A} := (L_2 D_2)^* D_1 (L_2 D_2) = M.$$

Such a modification forces a different mapping for odd and even steps and employs square roots.

The advantage of Theorem 2.1 over the explanation (1.1) mentioned in section 1 is that it reveals explicitly in (2.2) and (2.3) how the triangular factors L_1 and $L_2 D_2$ from LR yield the triangular factorization of Q from QR.

Remark 3. The QR transform does not require that A permit triangular factorization. In fact $\hat{A}$ cannot be derived from two steps of LR when, and only when, the orthogonal factor Q does not permit factorization as

$$Q = L_1 D_2^{-1} (D_2 L_2^{-*} D_2^{-1}).$$

In many cases, but not all, a well-chosen symmetric permutation $A \to \Pi A \Pi^t$ will give rise to a new Q that permits triangular factorization.

3. The non-Hermitian case. For general matrices the LR transform preserves band structure while the QR transform destroys the upper bandwidth. So the two procedures are not equivalent. Nevertheless it is legitimate to ask whether the QR transform can be represented in an alternative form related to triangular factorization.

The answer is yes. The key to extending the result of the previous section is to factor the given matrix B with a congruence transformation

$$B = FCF^*.$$

This appears to be a strange representation of a non-Hermitian matrix.

Suppose B permits triangular factorization

$$B = L_1 D_1 U_1.$$

Rewrite this as

$$B = L_1 (D_1 U_1 L_1^{-*}) L_1^*$$

and note that the middle factor is upper triangular instead of diagonal. Define, as earlier,

$$\overset{\circ}{B} = (D_1 U_1 L_1^{-*})(L_1^* L_1),$$

and use the Cholesky factorization

$$L_1^* L_1 = (L_2 D_2)(L_2 D_2)^*$$

to define

$$\begin{aligned} \overset{\circ\circ}{B} &= D_2 L_2^* (D_1 U_1 L_1^{-*}) L_2 D_2 \\ &= (D_2 L_2^* L_1^{-1})(L_1 D_1 U_1)(L_1^{-*} L_2 D_2) \\ &= Q^* B Q, \qquad \text{using (2.2).} \end{aligned}$$

Moreover,

$$\begin{aligned} B &= L_1 D_1 U_1 \\ &= (L_1 L_2^{-*} D_2^{-1})(D_2 L_2^* D_1 U_1) \\ &= Q\,\mathrm{sign}(D_1)(\overline{\mathrm{sign}(D_1)} D_2 L_2^* D_1 U_1) \end{aligned}$$

is the QR factorization of B.

Now, in general, $\mathrm{sign}(D_1) = \mathrm{diag}(\exp(i\varphi_1), \ldots, \exp(i\varphi_n))$.

Another way to interpret these expressions is to observe that, ignoring diagonal unitary matrices, the Q factor of B is the Q factor of its lower triangular factor L_1.

Note that $\overset{\circ}{B} = L_1^{-1} B L_1$, $\overset{\circ\circ}{B} = (D_2 L_2^*)\,\overset{\circ}{B}\,(D_2 L_2^*)^{-1}$, and so the nice unitary matrix Q is again split into its two triangular factors L_1 and $(D_2 L_2^*)^{-1}$.

Acknowledgments. The author would like to express his gratitude to Professors E.-X. Jiang, V. Mehrmann, B. Parlett, and D. Watkins for their encouragement and suggestions. He also thanks the referees for their valuable comments and criticism. Special thanks should be given to one of the referees who so kindly helped the author revise the manuscript.

REFERENCES

[1] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
[2] J. G. F. Francis, *The QR transformation: A unitary analogue to the LR transformation*, Parts I and II, Comput. J., 4 (1961), pp. 265–272 and pp. 332–345.
[3] K. Gates and W. B. Gragg, *Notes on TQR algorithms*, J. Comput. Appl. Math., 86 (1997), pp. 195–203.
[4] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[5] V. N. Kublanovskaya, *On some algorithms for the solution of the complete eigenvalue problem*, U.S.S.R. Comput. Math. and Math. Phys., 3 (1961), pp. 637–657.
[6] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[7] H. Rutishauser, *Solution of eigenvalue problems with the LR transformation,* Nat. Bur. Standards Appl. Math. Ser., 49 (1958), pp. 47–81.
[8] D. S. Watkins, *Understanding the QR algorithm*, SIAM Rev., 24 (1982), pp. 427–440.
[9] D. S. Watkins, *Fundamentals of Matrix Computations*, John Wiley, New York, 1991.
[10] D. S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 556–563, April 1998

018

COMPLEMENTATION IN STOCHASTIC MATRICES AND THE GTH ALGORITHM*

E. SENETA†

Abstract. The Grassmann, Taksar, and Heyman (GTH) algorithm for the computation of the stationary distribution of a finite stochastic matrix is shown to apply for the general case when there is a unique stationary distribution. The approach is elementary and matrix based, with probabilistic arguments avoided, to give insight into the essential structural properties. A byproduct is a necessary and sufficient determinantal condition for regularity of a stochastic matrix.

Key words. Markov chains, regular, stationary distribution, partitioned inversion, stochastic complement, Bartlett identity, determinantal conditions

AMS subject classifications. 60J10, 15A51, 60–04, 65U05

PII. S0895479896310172

1. Introduction.

Suppose for the moment $P = \{p_{ij}\},\ i,j = 0,1,\ldots,N$, is an irreducible aperiodic stochastic matrix ($p_{ij} \geq 0,\ \sum_{j=0}^{N} p_{ij} = 1$), so there is a unique stationary vector $\mathbf{p} = \{p_i\},\ \mathbf{p} > \mathbf{0}\,(\mathbf{p}^T P = \mathbf{p}^T,\ \mathbf{p}^T\mathbf{1} = 1)$.

Writing $A = P - I$, Grassmann, Taksar, and Heyman (GTH) [1] have shown how to construct an elegant and economical algorithm for the computation of $\mathbf{p}$ by applying, initially, the steps of Gaussian elimination to the matrix A. That is, in $\mathbf{p}^T A = \mathbf{0}^T$ considered as equations with labels $0,1,\ldots,N$, first solve for p_N in the Nth equation, and eliminate p_N from the other equations. Then solve equation $N-1$ for p_{N-1}, and eliminate p_{N-1} from all other equations, except for equation N. Continue in this way until the equation 1 is solved for p_1 in terms of p_0.

Equation 0 is redundant and can be omitted. Then if a_{ij}^n are the values obtained before solving for p_n, then (with δ_{ij} the Kronecker delta)

$$
\begin{aligned}
a_{ij}^N &= p_{ij} - \delta_{ij} \\
a_{in}^{n-1} &= -a_{in}^n / a_{nn}^n\,, && 0 \leq i < n, \\
a_{ij}^{n-1} &= a_{ij}^n + a_{nj}^n\, a_{in}^{n-1}\,, && 0 \leq i < n, \quad 0 \leq j < n, \\
a_{ij}^{n-1} &= a_{ij}^n\,, && i \geq n \text{ or } j > n,
\end{aligned} \tag{1}
$$

for $n = N,\ldots,1$. A crucial element of this argument is that the pivotal element a_{nn}^n is never zero. GTH show this using a sophisticated probabilistic argument, which also shows that

$$
a_{nn}^n = -\sum_{j=0}^{n-1} a_{nj}^n \tag{2}
$$

*Received by the editors October 7, 1996; accepted for publication (in revised form) by G. P. Styan April 17, 1997.

http://www.siam.org/journals/simax/19-2/31017.html

†School of Mathematics and Statistics, F07, University of Sydney, NSW 2006, Australia (seneta_e@maths.su.oz.au). This work was done in part while the author was a Visiting Professor at the Department of Statistics, University of Chicago, Chicago, IL, 1994.

so that, from (1),

$$a_{in}^{n-1} = a_{in}^{n} \Big/ \sum_{j=0}^{n-1} a_{nj}^{n}\,, \quad 0 \le i < n\,.$$

The above forms Item 1 of the following simple algorithm [1, p. 1112] (see also [2]).

Item 1. For $n = N, N-1, \ldots, 1$, do the following:

$$\begin{aligned} \text{Let } S &= \sum_{j=0}^{n-1} a_{nj}. \\ \text{Let } a_{in} &= a_{in}/S\,, \quad i < n. \\ \text{Let } a_{ij} &= a_{ij} + a_{in}\, a_{nj}\,, \quad i, j < n. \end{aligned}$$

Item 2. Let TOT $= 1,\ r_0 = 1$.

Item 3. For $j = 1, 2, \ldots, N$, do the following:

$$\begin{aligned} \text{Let } r_j &= \sum_{k=0}^{j-1} r_k\, a_{kj}. \\ \text{Let } TOT &= TOT + r_j. \end{aligned}$$

Item 4. Let $p_j = r_j/TOT\,,\ j = 0, 1, \ldots, N$.

Note that the line under Item 3 may also be written $r_j = a_{0j} + \sum_{k=1}^{j-1} r_k\, a_{kj}$.

Items 2, 3, and 4 of the algorithm arise in GTH from the same sophisticated probabilistic argument, and lead to a probabilistic interpretation of the quantities $r_j = p_j/p_0,\ 1 \le j \le N$.

The algorithm, for this setting, may well be close to optimal. The number of operations, on account of underlying Gaussian elimination, is asymptotically optimal at order $2N^3/3$. The a_{ij} with $i \ne j$ are always nonnegative, and no subtractions are used. According to GTH, it is therefore still effective for problems of size $N = 1000$. Stewart [12, p. 68] states that although this algorithm remains unanalyzed from the numerical standpoint, he believes it is stable and should be used routinely. A numerical analysis by O'Cinneide [9] has appeared in the same year.

We examine the matrix underpinnings of the algorithm, essentially in sections 3 and 4, without probabilistic intervention in the form of regenerative notions used in GTH. This is done in sections 2–4 under the *more general assumption* that the stochastic matrix $P = \{p_{ij}\}, \quad i, j = 0, 1, \ldots, N$ contains a single irreducible set of indices. We call such a stochastic matrix *regular*. This assumption is necessary and sufficient for there to be a unique stationary distribution vector $\mathbf{p} \ge \mathbf{0}$. (The entries of this vector corresponding to the irreducible set of indices are positive, and are zero otherwise.)

We also *assume* in regard to the algorithm, without loss of generality, that the labeling of the indices is such that zero is in the irreducible set.

We note Heyman [4] who shows implicitly that if P is irreducible (so $\mathbf{p} > \mathbf{0}$) and periodic, and the matrix P is first rewritten in the usual cyclic-block structure, then the algorithm works (and computational economies result). Earlier Heyman [3, p. 227] had asserted without proof that the algorithm can be made to work in the regular case under a less general relabeling.

Our treatment covers the most general structure of P giving unique stationary distribution, and gives some insight into what specific matrix features of the core irreducibility assumption make the procedure work. This leads to some understanding of how the setting may be generalized away from the usual Markovian situations, and is in itself consistent with matrix theory approaches to finite Markov chains.

We shall use the convention (as above) that a lowercase boldface letter denotes a column vector. The vector $\mathbf{e}$ used in section 3 does *not* have the conventional meaning of a column of ones.

2. Partition and complementation: Background. Recall that a stochastic matrix P with a single irreducible set of indices may, with suitable permutation of indices, be written in the canonical form

$$P = \begin{bmatrix} P_1 & 0 \\ R & Q \end{bmatrix}, \quad \text{so} \quad P^k = \begin{bmatrix} P_1^k & 0 \\ R_k & Q^k \end{bmatrix}, \tag{3}$$

where P_1 contains the matrix entries within the irreducible class of indices, and Q refers to indices (if any) outside this class, where $Q^k \to 0 \quad \text{as} \quad k \to \infty$. The fact that there is a solution $\mathbf{p}$ to $\mathbf{p}^T P = \mathbf{p}^T, \mathbf{p} \neq \mathbf{0}$, and is unique to constant multiples and may be taken as a multiple of an elementwise nonnegative vector (with the positive entries corresponding to the irreducible set) follows, without probabilistic reasoning, from the fact that $Q^k \to 0$ and the Perron–Frobenius theory [11]. The norming $\mathbf{p}^T \mathbf{1} = 1$ then specifies $\mathbf{p}(\geq \mathbf{0})$ uniquely as the stationary vector.

Suppose now the index 0 of the set of indices to be a member of the irreducible set. Writing $P^k = \{p_{ij}^{(k)}\}$, it is seen from (3), since P_1 is irreducible and since with increasing k the row sums of R_k tend to unity, that for any $i \epsilon \{0, 1, \ldots, N\}$ there is a $k \equiv k(i)$ such that $p_{i0}^{(k)} > 0$. Thus, for any subset J of $\{0, 1, 2, \ldots, N\}$ containing the index 0, for this same $k \equiv k(i)$

$$\sum_{j \epsilon J} p_{ij}^{(k)} > 0\,. \tag{4}$$

The properties discussed above are clearly not dependent on the reordering of the indices $\{0, 1, 2, \ldots, N\}$ to give the canonical form (3). Thus, we assume $P = \{p_{ij}\},\ i, j = 0, 1, \ldots, N$ to be in some arbitrary initial given form, *with the proviso* that the first index, 0, is in the irreducible set. Consider the regular matrix P now partitioned as

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix},$$

where P_{11} is $(n \times n), n = 1, \ldots, N$, and $\mathbf{p}^T = \{\mathbf{p}_1^T\,,\ \mathbf{p}_2^T\}$ is partitioned accordingly.

LEMMA 2.1. *$I - P_{22}$ is nonsingular and $(I - P_{22})^{-1} = \sum_{k=0}^{\infty} P_{22}^k$ elementwise.*

Proof. Consider the stochastic matrix $\tilde{P} = \{\tilde{p}_{ij}\},\ i, j = 0, 1, \ldots, N - n + 1$,

$$\tilde{P} = \begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{p}_0 & P_{22} \end{bmatrix}.$$

Then for $i \epsilon \{1, 2, \ldots, N - n + 1\}$, using (4) and regarding J there as the set of indices which here are compressed into the zero index, there is a $k \equiv k(i)$ such that

$$\tilde{p}_{i0}^{(k)} > 0, \quad \text{so} \quad \sum_{j=1}^{N-n+1} \tilde{p}_{ij}^{(k)} < 1,$$

where $\tilde{P}^k = \{\tilde{p}_{ij}^{(k)}\}$. Then following the proof of Seneta [11, Theorem 4.3], with $I = \{1, 2, \ldots, N-n+1\}$ and $Q = P_{22}$ we find $P_{22}^k \to 0$. The conclusion follows from Lemma B.1 of Appendix B in Seneta [11]. □

LEMMA 2.2. *The $n \times n$ matrix*

$$S_{11} = P_{11} + P_{12}(I - P_{22})^{-1}P_{21} \tag{5}$$

is stochastic and regular, and the corresponding stationary distribution vector is $\mathbf{p}_1/\mathbf{p}_1^T\mathbf{1}$.

Proof. The various components of the lemma are well known for irreducible P (see, for example, Kemeny and Snell [6, 6.1.1] and Meyer [8, Theorems 2.1 and 2.2]). To recapitulate briefly, continuing to avoid any probabilistic argument, we have that

$$\begin{aligned} \mathbf{p}_1^T P_{11} + \mathbf{p}_2^T P_{21} &= \mathbf{p}_1^T, \\ \mathbf{p}_1^T P_{12} + \mathbf{p}_2^T P_{22} &= \mathbf{p}_2^T, \end{aligned}$$

where the first entry of $\mathbf{p}_1^T$ is positive, since zero is in the irreducible set of P.

From the second of these equations and Lemma 2.1, $\mathbf{p}_2^T = \mathbf{p}_1^T P_{12}(I - P_{22})^{-1}$ so that, from the first equation, $\mathbf{p}_1^T(P_{11} + P_{12}(I - P_{22})^{-1} P_{21}) = \mathbf{p}_1^T$. Again by Lemma 2.1, the matrix S_{11} is elementwise nonnegative, and stochastic since in $P_{11}\mathbf{1} + P_{12}(I - P_{22})^{-1}P_{21}\mathbf{1}$ we have that $P_{21}\mathbf{1} = (I - P_{22})\mathbf{1}$. If S_{11} does not have a unique stationary distribution, it is possible to find a nonzero solution $\mathbf{x}_1$ to $\mathbf{x}_1^T S_{11} = \mathbf{x}_1^T$ which is nonunique to constant multiples. Putting $\mathbf{x}_2^T = \mathbf{x}_1^T P_{12}(I - P_{22})^{-1}$, the solution $\mathbf{x}$ to $\mathbf{x}^T P = \mathbf{x}^T$, where $\mathbf{x}^T = (\mathbf{x}_1^T, \mathbf{x}_2^T)$, would have this property, which is a contradiction to the fact that regular P has unique stationary distribution. □

The matrix S_{11} corresponding to P is called the *stochastic complement* of P_{11} in P. It is the central idea behind the next section, in showing that the pivotal elements are all nonzero. A connection between the ideas behind the GTH algorithm and stochastic complementation is touched on by Heyman [5].

3. Partitioned inversion and coefficient matrices. In what follows we write $S(n-1)$ in place of S_{11} of (5) and

$$P = \begin{bmatrix} P_{11} & \mathbf{a} & B \\ \mathbf{b}^T & c & \mathbf{e}^T \\ C & \mathbf{d} & D \end{bmatrix},$$

where P_{11} is $n \times n$, corresponding to indices $0, 1, \ldots, n-1$. Thus, for $N \geq n \geq 1$, keeping in mind Lemmas 2.1 and 2.2,

$$S(n-1) = P_{11} + (\mathbf{a}\ B)\left[I - \begin{bmatrix} c & \mathbf{e}^T \\ \mathbf{d} & D \end{bmatrix}\right]^{-1} \begin{bmatrix} \mathbf{b}^T \\ C \end{bmatrix}. \tag{6}$$

$$S(n) = \begin{bmatrix} P_{11} & \mathbf{a} \\ \mathbf{b}^T & c \end{bmatrix} + \begin{bmatrix} \mathbf{B} \\ \mathbf{e}^T \end{bmatrix}(I - D)^{-1}(C\ \mathbf{d}) \tag{7}$$

with $S(N) = P$. Write $\mathbf{p}^T = (\mathbf{p}_1^T(n-1)\,,\ \mathbf{p}_2^T(n-1)) = (\mathbf{p}_1^T(n)\,, \mathbf{p}_2^T(n))$, the partitions according with $S(n-1)$ and $S(n)$, respectively.

By partitioned inversion

$$\left[I - \begin{bmatrix} c & \mathbf{e}^T \\ \mathbf{d} & D \end{bmatrix}\right]^{-1} = \begin{bmatrix} k & (1-c)^{-1}\mathbf{e}^T K \\ k(I-D)^{-1}\,\mathbf{d} & K \end{bmatrix}, \tag{8}$$

where $k = (1 - c - \mathbf{e}^T(I - D)^{-1}\mathbf{d})^{-1}$ and

$$K = (I - D - (1 - c)^{-1}\mathbf{d}\,\mathbf{e}^T)^{-1} = (I - D)^{-1} + k(I - D)^{-1}\mathbf{d}\,\mathbf{e}^T(I - D)^{-1}, \tag{9}$$

the last expression following from the Bartlett (Sherman–Morrison) identity, since $(I - D)^{-1}$ exists and

$$1 - c - \mathbf{e}^T(I - D)^{-1}\mathbf{d} \neq 0 \tag{10}$$

from the *stochasticity and regularity* of $S(n)$, since otherwise existence of more than one irreducible set would be implied (recall that zero is already in one irreducible set).

Next, write $\mathbf{p}_1^T(n) = (\mathbf{p}_1^T(n-1), p_n)$, and consider the first step of Gaussian elimination (that is, for p_n) in the system for $1 \leq n \leq N$

$$\mathbf{p}_1^T(n)\,A(n) = \mathbf{0}^T, \text{ where } A(n) = S(n) - I. \tag{11}$$

From (7) we find that

$$p_n(1 - c - \mathbf{e}^T(I - D)^{-1}\,\mathbf{d}) = \mathbf{p}_1^T\,(n-1)\{\mathbf{a} + B(I - D)^{-1}\mathbf{d}\}, \tag{12}$$

where, from (10) the coefficient of p_n, viz., k^{-1}, is not zero. Thus, for $1 \leq n \leq N$

$$p_n = \mathbf{p}_1^T\,(n-1)\{k(\mathbf{a} + B(I - D)^{-1}\mathbf{d})\}. \tag{13}$$

Substituting (13) into the remaining equations of (11) we obtain

$$\mathbf{p}_1^T(n-1)\{P_{11} - I + B(I-D)^{-1}\,C + k\{\mathbf{a} + B(I-D)^{-1}\mathbf{d}\}\,\{\mathbf{b}^T + \mathbf{e}^T(I-D)^{-1}\,C\}\} = \mathbf{0}^T. \tag{14}$$

We now prove that the matrix appearing on the left of (14) is in fact $A(n-1) = S(n-1) - I$. This enables us to conclude that the pivotal element of $A(n-1)$ analogous to that of $A(n)$ is likewise nonzero in the manner of (10) for $A(n)$, since $A(n-1)$ has parallel structure to $A(n)$, $N \geq n \geq 2$. The fact that (2) holds for each n, $N \geq n \geq 1$, is now merely a consequence of the stochasticity of each $S(n)$, which follows from Lemma 2.2.

LEMMA 3.1. *The coefficient matrix in* (14) *is* $A(n-1) = S(n-1) - I$.

Proof. From (6) and (8)

$$S(n-1) - I = P_{11} - I + k\{\mathbf{a} + B(I-D)^{-1}\,\mathbf{d}\}\,\mathbf{b}^T + (1-c)^{-1}\,\mathbf{a}\,\mathbf{e}^T\,KC + BKC\,. \tag{15}$$

Now, from (9)

$$KC = (I - D)^{-1}\,C + k(I - D)^{-1}\,\mathbf{d}\,\mathbf{e}^T\,(I - D)^{-1}\,C\,,$$

so (15) becomes

$$\begin{aligned} A(n-1) = {} & P_{11} - I + B(I - D)^{-1}\,C + k\{\mathbf{a} + B(I - D)^{-1}\,\mathbf{d}\}\,\mathbf{b}^T \\ & + (1-c)^{-1}\mathbf{a}\,\mathbf{e}^T\,(I - D)^{-1}\,C + (1-c)^{-1}\,\mathbf{a}\,\mathbf{e}^T\,k(I - D)^{-1}\,\mathbf{d}\,\mathbf{e}^T(I - D)^{-1}\,C \\ & + k\,B\,(I - D)^{-1}\,\mathbf{d}\,\mathbf{e}^T\,(I - D)^{-1}\,C. \end{aligned} \tag{16}$$

Focus on the two penultimate terms of (16): their sum simplifies to

$$(1 - c)^{-1}\,\mathbf{a}\{1 + k\,\mathbf{e}^T(I - D)^{-1}\,\mathbf{d}\}\,\mathbf{e}^T(I - D)^{-1}\,C$$

$$= (1 - c)^{-1}\,(1 + k\,\mathbf{e}^T(I - D)^{-1}\,\mathbf{d})\,\mathbf{a}\,\mathbf{e}^T(I - D)^{-1}\,C\,. \tag{17}$$

Now, since $k^{-1} = (1 - c - \mathbf{e}^T(I - D)^{-1}\,\mathbf{d})$ we see that

$$k(1 - c) = 1 + k\,\mathbf{e}^T(I - D)^{-1}\,\mathbf{d},$$

so

$$(1 - c)^{-1}\,(1 + k\,\mathbf{e}^T(I - D)^{-1}\,\mathbf{d}) = k. \tag{18}$$

Combining (18), (17), and (16), and comparing with (14) completes the proof. □

4. The probability-normed solution. We now justify Items 2, 3, 4 of the algorithm restated in section 1 from GTH without probabilistic interpretation of p_1/p_0. From the structure of the coefficient matrices as demonstrated in section 3, at the last stage we are able to solve for p_1 in terms of p_0. Let us put $\tilde{p}_0 = 1$, so that the corresponding $p_1 = \tilde{p}_1$ may be obtained by back substitution, and so on to obtain all elements of $\tilde{p} = \{\tilde{p}_i\}\,,\ i = 0, 1, \ldots, N$.

Indeed, the key back substitution step may be expressed as

$$\tilde{p}_n = \tilde{\mathbf{p}}_1^T(n-1)\{\mathbf{a} + B(I - D)^{-1}\mathbf{d}\}/(1 - c - \mathbf{e}^T(I - D)^{-1}\,\mathbf{d}) \tag{19}$$

from (13), keeping in mind (10), and having written $\tilde{\mathbf{p}}_1^T(n) = (\tilde{\mathbf{p}}_1^T(n-1), \tilde{p}_n)$. The quantity "$S$" in the algorithm at this stage is just $(1 - c - \mathbf{e}^T(I - D)^{-1}\,\mathbf{d})$, and our equation (19) is equivalent to equations (20)–(21) of GTH, and $\tilde{p}_n$ to quantity r_n in the algorithm.

By the Perron–Frobenius theorem for P, we have that $\tilde{p}_j = \text{const.}\,p_j\,,\ j = 0, 1, \ldots, N$, and since $\sum_{j=0}^{N} p_j = 1$, we see that $\sum_{j=0}^{N} \tilde{p}_j = \text{const.}$, so $p_j = \tilde{p}_j / \sum_{i=0}^{N} \tilde{p}_i = r_j/TOT$ as in Item 4 of the algorithm.

Incidentally, from the uniqueness to constant multiples of the Perron–Frobenius eigenvector, it follows that $\tilde{p}_i = p_i/p_0\,,\ i = 0, 1, \ldots, N$, so that

$$\sum_{i=0}^{N} \tilde{p}_i = 1 + (1 - p_0)/p_0 = 1/p_0\,,$$

so that

$$p_j = p_0\,\tilde{p}_j = p_0\,r_j\,, \quad j = 1, \ldots, N,$$

which is the probabilistically deduced starting point for GTH's justification of final Items 2, 3, 4 of the algorithm.

5. Essence. In this concluding section we review the key matrix features of the preceding to show that the algorithm will work for certain matrices P which may not have the structure of either a probability transition matrix or a Markov intensity matrix. An example is the matrix

$$P = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & 2/3 & 2/3 & -1/3 \\ 1/3 & 0 & 4/3 & -2/3 \\ -1/3 & 1/3 & -1/3 & 4/3 \end{bmatrix}.$$

More generally, let $P = \{p_{ij}\}, \quad i, j = 0, 1, \ldots, N$, be a matrix of real entries and $A = P - I$. Consider P written in partitioned form as in section 3, and *assume*

$I - D$ is nonsingular for each $(N + 1 - n) \times (N + 1 - n)$ matrix D, $\quad n = 1, 2, \ldots, N$, which implies that the bottom right-hand corner entry of P is not unity. Successive $(I - D)^{-1}$ are then defined by (8) and (9) providing it is further *assumed* (10) holds for each before defining the next, so that each pivotal element is nonzero.

$S(n-1)$ and $S(n)$ are now defined by (6) and (7).

We notice that if an $(n \times 1)$ $\mathbf{p}_1(n-1)$ satisfies $\mathbf{p}_1^T(n-1)\,A(n-1) = \mathbf{0}^T$, where $A(n-1) = S(n-1) - I$, and we define p_n by (13), then by reversing the argument of Lemma 3.1 and (14), we find that $\mathbf{p}_1^T(n)\,A(n) = \mathbf{0}^T$, where $\mathbf{p}_1^T(n) = (\mathbf{p}_1^T(n-1), p_n)$, and $A(n) = S(n) - I$. Thus, if we define $\tilde{p}_0 = 1$ as in section 4, and proceed through the back substitution steps (19) to produce $\tilde{\mathbf{p}} = \{\tilde{p}_j\}$, $j = 0, 1, \ldots, N$, this vector is a solution $\mathbf{p}, \mathbf{p} \neq \mathbf{0}$, to $\mathbf{p}^T A = \mathbf{0}^T$. Since $(D - I)$ is nonsingular in particular when it is of dimension $(N \times N)$, by one of the assumptions in this section it follows that $P - I$ has the last N rows of its $(N + 1)$ rows linearly independent. It is, however, singular, since we have just shown that there is a solution $\mathbf{p} \neq \mathbf{0}$ to $\mathbf{p}^T(P - I) = \mathbf{0}^T$. Thus, the 0th row of $P - I$ is a linear combination of the last N rows. Thus, any solution $\mathbf{p} = \{p_i\} \neq \mathbf{0}$ to $\mathbf{p}^T A = \mathbf{0}^T$ must have $p_0 \neq 0$ and must be of form $p_0\tilde{\mathbf{p}}$; that is, $\tilde{p}_i = p_i/p_0$, $i = 0, 1, \ldots, N$.

In summary: (1) the conditions which P above is supposed to satisfy imply that there is a nonzero solution $\mathbf{p}$ to $\mathbf{p}^T P = \mathbf{p}^T$, with 0th position entry of $\mathbf{p}$ not zero, and this solution is unique to constant multiples (thus, one is an eigenvalue of P and $\mathbf{p}$ is a corresponding left eigenvector, unique to constant multiples). (2) The solution may be obtained by Gaussian elimination with all pivots nonzero, and simple back substitution, after setting $\tilde{p}_0 = 1$.

We have retained the "equal row sums" property possessed by stochastic matrices in our numerical example above. In this example the $\{\tilde{p}_i\}$, $i = 0, 1, 2, 3$, is $(1, -2, 0, -3)$. The conditions would have continued to apply to the example if we had made it (5×5) by adding a final (1×4) row of zeros, a final 1×4 column of zeros, and a $(5, 5)$ element $1/2$, say to remove resemblance to irreducibility and stochasticity. The vector $\{\tilde{p}_i\}$ would then have a fifth entry equal to zero.

The conditions imposed on P in this section have perhaps some independent interest in that they resemble the determinant conditions imposed by Markov [7]—see Schneider [10] for an analysis—on a finite *stochastic* matrix in place of irreducibility. Indeed, if a stochastic matrix P has more than one irreducible (i.e., closed) class of indices, the condition that $I - D$ is nonsingular for each $(N + 1 - n) \times (N + 1 - n)$ matrix D, $n = 1, 2, \ldots, N$, is broken. To see this, note that the subset of indices corresponding to one of the irreducible classes must be within the set $\{1, 2, \ldots, N\}$ of the index set $\{0, 1, 2, \ldots, N\}$ and one of the $I - D$'s will contain these indices, and its determinant will be zero. This may be seen by considering a simultaneous permutation of rows and columns of this $I - D$ to obtain a canonical form of D which has an isolated stochastic matrix (corresponding to the irreducible class) on the diagonal. Thus, $\det(I - D) = 0$, since a simultaneous permutation of rows and columns is a similarity transformation.

The reader will notice that the condition is also broken if P has only one irreducible set of indices, if index 0 is not a member of this set.

Thus, regularity of a stochastic P is guaranteed if each index in turn is considered as the zero index, and the condition that each $(N+1-n) \times (N+1-n)$ matrix D, $n = 1, 2, \ldots, N$, is nonsingular is satisfied for one such choice. This is thus a necessary and sufficient determinantal condition for *regularity*, a more fundamental structural property of stochastic P in Markov chain applications than irreducibility. Of course such determinantal conditions are now of theoretical interest only.

REFERENCES

[1] W. K. Grassmann, M. I. Taksar, and D. P. Heyman, *Regenerative analysis and steady state distributions for Markov chains*, Oper. Res., 33 (1985), pp. 1107–1116.

[2] W. K. Grassmann, *Means and variances in Markov reward systems*, in Linear Algebra, Markov Chains, and Queueing Models, C. D. Meyer and R. J. Plemmons, eds., Springer, New York, 1993, pp. 193–204.

[3] D. P. Heyman, *Further comparisons of direct methods for computing stationary distributions of Markov chains*, SIAM J. Alg. Disc. Meth., 8 (1987), pp. 226–232.

[4] D. P. Heyman, *A direct algorithm for computing the stationary distribution of a p-cyclic Markov chain*, in Linear Algebra, Markov Chains, and Queueing Models, C. D. Meyer and R. J. Plemmons, eds., Springer, New York, 1993, pp. 205–209.

[5] D. P. Heyman, *A decomposition theorem for infinite stochastic matrices*, J. Appl. Probab., 32 (1995), pp. 893–901.

[6] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Van Nostrand, Princeton, NJ, 1960.

[7] A. A. Markov, *Generalization of limit theorems of the probability calculus to sums of chain dependent quantities*, 1908, in Izbrannie Trudy (Selected Works), Izd. AN SSSR, Moscow, 1951, pp. 365–397 (in Russian).

[8] C. D. Meyer, *Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems*, SIAM Rev., 31 (1989), pp. 240–272.

[9] C. A. O'Cinneide, *Entrywise perturbation theory and error analysis for Markov chains*, Numer. Math., 65 (1993), pp. 109–120.

[10] H. Schneider, *The concepts of irreducibility and full indecomposability of a matrix in the works of Frobenius, König and Markov*, Linear Algebra Appl., 18 (1977), pp. 139–162.

[11] E. Seneta, *Non-Negative Matrices and Markov Chains*, Springer, New York, 1981.

[12] G. W. Stewart *Gaussian elimination, perturbation theory and Markov chains*, in Linear Algebra, Markov Chains, and Queueing Models, C. D. Meyer and R. J. Plemmons, eds., Springer, New York, 1993, pp. 59–69.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 2, pp. 564–581, April 1998

019

CONDITION NUMBERS OF RANDOM TRIANGULAR MATRICES*

D. VISWANATH† AND L. N. TREFETHEN‡

Abstract. Let L_n be a lower triangular matrix of dimension n each of whose nonzero entries is an independent $N(0,1)$ variable, i.e., a random normal variable of mean 0 and variance 1. It is shown that κ_n, the 2-norm condition number of L_n, satisfies

$$\sqrt[n]{\kappa_n} \to 2 \quad \textit{almost surely}$$

as $n \to \infty$. This exponential growth of κ_n with n is in striking contrast to the linear growth of the condition numbers of random *dense* matrices with n that is already known. This phenomenon is not due to small entries on the diagonal (i.e., small eigenvalues) of L_n. Indeed, it is shown that a lower triangular matrix of dimension n whose diagonal entries are fixed at 1 with the subdiagonal entries taken as independent $N(0,1)$ variables is also exponentially ill conditioned with the 2-norm condition number κ_n of such a matrix satisfying

$$\sqrt[n]{\kappa_n} \to 1.305683410\ldots \quad \textit{almost surely}$$

as $n \to \infty$. A similar pair of results about complex random triangular matrices is established. The results for real triangular matrices are generalized to triangular matrices with entries from any symmetric, strictly stable distribution.

Key words. random triangular matrices, matrix condition numbers, exponentially nonnormal matrices, strong limit theorems

AMS subject classifications. 15A52, 15A12, 65F35, 60F15

PII. S0895479896312869

1. Introduction.

Random dense matrices are well conditioned. Edelman has shown that if each of the n^2 entries of a matrix of dimension n is an independent $N(0,1)$ variable, the probability density function (PDF) of κ_n/n, where κ_n is the 2-norm condition number of such a matrix, converges pointwise to the function

$$\frac{2x+4}{x^3}\exp(-2x^{-1}-2x^{-2})$$

as $n \to \infty$ [5]. Since the distribution of κ_n/n is independent of n in the limit $n \to \infty$, we can say that the condition numbers of random dense matrices grow only linearly with n. Using this PDF, it can be shown, for example, that $E(\log(\kappa_n)) = \log(n) + 1.537\ldots + o(1)$ [5].

In striking contrast, the condition number of a random lower triangular matrix L_n, a matrix of dimension n all of whose diagonal and subdiagonal entries are independent $N(0,1)$ variables, grows exponentially with n. If κ_n is the 2-norm condition number of L_n (defined as $\|L_n\|_2\|L_n^{-1}\|_2$), we show that

$$\sqrt[n]{\kappa_n} \to 2 \quad \textit{almost surely}$$

as $n \to \infty$ (Theorem 4.3). Figure 1.1a illustrates this result.

*Received by the editors December 2, 1996; accepted for publication by A. Edelman April 21, 1997. This work was supported by NSF grant DMS-9500975CS and DOE grant DE-FG02-94ER25199.
http://www.siam.org/journals/simax/19-2/31286.html
†Department of Computer Science, Cornell University, Ithaca, NY 14853 (divakar@cs.cornell.edu).
‡Computing Laboratory, Oxford University, Oxford, 0X1 3QD, UK (lnt@comlab.ox.uc.uk).

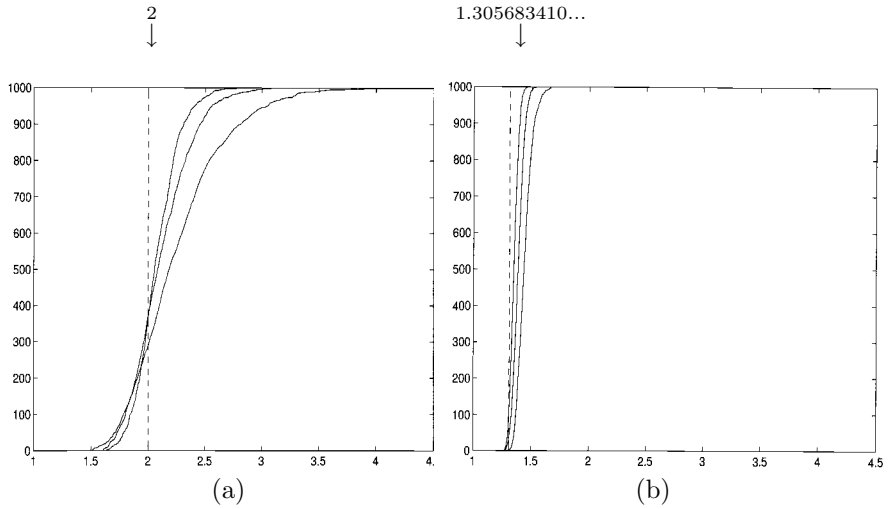


FIG. 1.1. *Empirical cumulative density functions of* $\sqrt[n]{\kappa_n}$*, for triangular and unit triangular matrices, respectively, with* $n = 25, 50, 100$ *obtained from* 1000 *random matrices for each* n*. The random entries are* $N(0,1)$ *variables. The higher values of* n *correspond to the steeper curves. In the limit* $n \to \infty$*, the cumulative density functions converge to Heaviside step functions with jumps at the dashed lines.*

The matrices that arise in the experiments reported in Figure 1.1 are so ill conditioned that the standard, normwise stable method of finding the condition number using the SVD [10] fails owing to rounding errors. The method used to generate the figures finds the inverse of the triangular matrix explicitly using the standard algorithm for triangular inversion, and then computes the norms of the matrix and its inverse independently. This works because the computation of each column of the inverse by the standard triangular inversion algorithm is componentwise backward stable [12].

The exponential growth of κ_n with n is not due to small entries on the diagonal since the probability of a diagonal entry being exponentially small is also exponentially small. For a further demonstration that the diagonal entries do not cause the exponential growth in κ_n, we consider condition numbers of *unit* triangular matrices, i.e., triangular matrices with ones on the diagonal. If κ_n is the condition number of a unit lower triangular matrix of dimension n with subdiagonal entries taken as independent $N(0,1)$ variables, then

$$\sqrt[n]{\kappa_n} \to 1.305683410\ldots \quad \textit{almost surely}$$

as $n \to \infty$ (Theorem 5.3). Obviously, in this case the ill conditioning has nothing to do with the diagonal entries (i.e., the eigenvalues) since they are all equal to 1. The relationship of the exponential ill conditioning of random unit triangular matrices to the stability of Gaussian elimination with partial pivoting is discussed in section 7.

We will use L_n to refer to triangular matrices of various kinds — real or complex, with or without a unit diagonal. But L_n always denotes a lower triangular matrix of dimension n. If the entries of L_n are random variables, they are assumed to be independent. Thus, if we merely say that L_n has entries from a certain distribution, those entries are not only identically distributed but also independent. Of course, only the nonzero entries of L_n are chosen according to that distribution. The condition number always refers to the 2-norm condition number. However, all our results concerning the limits $\lim_{n\to\infty} \sqrt[n]{\kappa_n}$ apply to all the L_p norms, $1 \le p \le \infty$, since

$n^{1/n} \to 1$ as $n \to \infty$ and the L_p norms differ by at most a factor of n. The 2-norm condition number of L_n, defined as $\|L_n\|_2 \|L_n^{-1}\|_2$, is denoted by κ_n. The context will make clear the distribution of the entries of L_n.

The analyses and discussions in this paper are phrased for lower, not upper, triangular matrices. However, all the theorems are true for upper triangular matrices as well, as is obvious from the fact that a matrix and its transpose have the same condition number.

We obtain similar results for triangular matrices with entries chosen from the complex normal distribution $\tilde{N}(0, \sigma^2)$. By $\tilde{N}(0, \sigma^2)$ we denote the complex normal distribution of mean 0 and variance σ^2 obtained by taking the real and imaginary parts as independent $N(0, \sigma^2/2)$ variables. Let L_n denote a triangular matrix with $\tilde{N}(0, \sigma^2)$ entries. Then

$$\sqrt[n]{\kappa_n} \to e^{1/2} \quad \textit{almost surely}$$

as $n \to \infty$ (Theorem 7.3). Since $e^{1/2} < 2$, triangular matrices with complex normal entries tend to have smaller condition numbers than triangular matrices with real normally distributed entries.

Similarly, let L_n denote a unit lower triangular matrix with $\tilde{N}(0, 1)$ subdiagonal entries. Then

$$\sqrt[n]{\kappa_n} \to 1.347395784\ldots \quad \textit{almost surely}$$

as $n \to \infty$ (Theorem 7.4). Thus, unit triangular matrices with complex normal entries tend to have slightly bigger condition numbers than unit triangular matrices with real normal entries.

Our results are similar in spirit to results obtained by Silverstein for random dense matrices [16]. Consider a matrix of dimension $n \times (yn)$, where $y \in [0, 1]$, each of whose $n^2 y$ entries is an independent $N(0, 1)$ variable. Denote its largest and smallest singular values by $\sigma_{\max}$ and $\sigma_{\min}$, respectively. It is shown in [16] that

$$\frac{\sigma_{\max}}{\sqrt{n}} \to 1 + \sqrt{y}, \quad \frac{\sigma_{\min}}{\sqrt{n}} \to 1 - \sqrt{y} \quad \textit{almost surely}$$

as $n \to \infty$. The complex analogues of these results can be found in [4]. The technique used in [16] is a beautiful combination of what is now known as the Golub–Kahan bidiagonalization step in computing the SVD with the Gerschgorin circle theorem and the Marčenko–Pastur semicircle law. The techniques used in this paper are more direct.

The exponential growth of $\kappa_n = \|L_n\|_2 \|L_n^{-1}\|_2$ is due to the second factor. We outline the approach for determining the rate of exponential growth of κ_n by assuming L_n triangular with $N(0, 1)$ entries. In section 2, we derive the joint probability density function (JPDF) for the entries in any column of L_n^{-1} (Proposition 2.1). If T_k is the 2-norm of column $n-k+1$ of L_n^{-1}, i.e., the column with k nonzero entries, both positive and negative moments of T_k are explicitly derived in section 3 (Lemma 3.2). These moments allow us to deduce that $\sqrt[n]{\kappa_n}$ converges to 2 almost surely (Theorem 4.3). A similar approach is used to determine the limit of $\sqrt[n]{\kappa_n}$ for L_n unit triangular with $N(0, \sigma^2)$ entries, triangular with $\tilde{N}(0, \sigma^2)$ entries, and unit triangular with $\tilde{N}(0, \sigma^2)$ entries (Theorems 5.3, 7.3, and 7.4, respectively).

The same approach is used more generally to determine the limit of $\sqrt[n]{\kappa_n}$ as $n \to \infty$ for L_n with entries drawn from any symmetric, strictly stable distribution (Theorems 8.4 and 8.6). These theorems are specialized to the Cauchy distribution, which is symmetric and strictly stable, in Theorems 8.5 and 8.7.

(a) (b)

FIG. 2.1. *Entries of* L_n^{-1} *on the same solid line in* (a) *have the same PDF. Sets of entries of* L_n^{-1} *in the boxes in* (b) *have the same JPDF.*

2. Inverse of a random triangular matrix. Consider the matrix

$$L_n = \begin{pmatrix} \alpha_{11} & & & \\ -\alpha_{21} & \alpha_{22} & & \\ \vdots & \vdots & \ddots & \\ -\alpha_{n1} & -\alpha_{n2} & \dots & \alpha_{nn} \end{pmatrix},$$

where each α_{ij} is an independent $N(0,1)$ variable. Then L_n^{-1} is also lower triangular. Denote the first k entries in the first column of L_n^{-1} by $t_1, \dots, t_k$. The t_i satisfy the following relations:

$$\begin{aligned} t_1 &= 1/\alpha_{11}, \\ t_2 &= (\alpha_{21}t_1)/\alpha_{22}, \\ t_3 &= (\alpha_{31}t_1 + \alpha_{32}t_2)/\alpha_{33}, \\ &\vdots \\ t_k &= (\alpha_{k1}t_1 + \cdots + \alpha_{k,k-1}t_{k-1})/\alpha_{kk}. \end{aligned} \tag{2.1}$$

This system of equations can be interpreted as a system of random recurrence relations. The first entry t_1 is the reciprocal of an $N(0,1)$ variable. The kth entry t_k is obtained by summing the previous entries $t_1, \dots, t_{k-1}$ with independent $N(0,1)$ variables as coefficients, and dividing that sum by an independent $N(0,1)$ variable.

Next, consider an arbitrary column of L_n^{-1} and denote the first k entries of that column from the diagonal downwards by $t_1, \dots, t_k$. The entries t_i satisfy random recurrence relations similar in form to (2.1), but the α_{ij} are a different block of entries in L_n for different columns. For example, any diagonal entry of L_n^{-1} is the reciprocal of an $N(0,1)$ variable; in particular, the kth diagonal entry is $1/\alpha_{kk}$.

These observations about triangular inversion can be represented pictorially.

Every entry of L_n^{-1} at a fixed distance from the diagonal has the same PDF. We may say that the matrix L_n^{-1}, like L_n, is "statistically Toeplitz." See Figure 2.1a. Moreover, if we consider the first k entries of a column of L_n^{-1} from the diagonal downwards, those k entries will have the same JPDF irrespective of the column. See Figure 2.1b. The different columns of L_n^{-1}, however, are by no means independent.

The description of triangular inversion above and later arguments are stated in terms of the columns of L_n^{-1}. However, rows and columns are indistinguishable in this problem; we could equally well have framed the analysis in terms of rows.

Denote the JPDF of t_i, $1 \le i \le k$, by $f_k = f_k(t_1, \dots, t_k)$. In the next proposition, a recursive formula for f_k is derived. For simplicity, we introduce the further notation

$T_k = \sqrt{t_1^2 + \cdots + t_k^2}$. Throughout this section, L_n is the random triangular matrix of dimension n with $N(0,1)$ entries.

PROPOSITION 2.1. *The JPDF $f_k = f_k(t_1, \ldots, t_k)$ satisfy the following recurrence:*

$$f_1 = \frac{\exp(-1/2t_1^2)}{\sqrt{2\pi} t_1^2}, \tag{2.2}$$

$$f_k = \frac{1}{\pi} \frac{T_{k-1}}{T_k^2} f_{k-1} \quad \textit{for } k > 1. \tag{2.3}$$

Proof. The t_k are defined by the random recurrence in (2.1).

The expression for f_1 is easy to get. If x is an $N(0,1)$ variable, its PDF is

$$\frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

The change of variable $x = 1/t_1$ gives (2.2).

To obtain the recursive expression (2.3) for f_k, consider the variable τ_k obtained by summing the variables $t_1, \ldots, t_{k-1}$ as $\sum_{i=1}^{k-1} \alpha_{ki} t_i$, where α_{ki} are independent $N(0,1)$ variables. For fixed values of $t_i, 1 \le i \le k-1$, the variable τ_k, being a sum of random normal variables, is itself a random normal variable of mean 0 and variance T_{k-1}^2. Therefore, the JPDF of τ_k and $t_1, \ldots, t_{k-1}$ is given by

$$\frac{1}{\sqrt{2\pi}} \frac{\exp(-\tau_k^2/2T_{k-1}^2)}{T_{k-1}} f_{k-1}.$$

By (2.1), the variable t_k can be obtained as τ_k/α, where α is an independent $N(0,1)$ variable. The JPDF of α, τ_k, and $t_1, \ldots, t_{k-1}$ is given by

$$\frac{1}{\sqrt{2\pi}} \exp(-\alpha^2/2) \frac{1}{\sqrt{2\pi}} \frac{\exp(-\tau_k^2/2T_{k-1}^2)}{T_{k-1}} f_{k-1}.$$

Changing the variable τ_k to $t_k = \tau_k/\alpha$ and integrating out α, we obtain

$$f_k = \frac{1}{\pi} \frac{T_{k-1}}{T_{k-1}^2 + t_k^2} f_{k-1} = \frac{1}{\pi} \frac{T_{k-1}}{T_k^2} f_{k-1},$$

i.e., f_k is given by (2.3). ∎

Note that the form of the recurrence for f_k in Proposition 2.1 mirrors the random recurrence (2.1) for obtaining t_k from the previous entries $t_1, \ldots, t_{k-1}$. In the following corollary, an explicit expression for f_k in terms of the t_i is stated.

COROLLARY 2.2. *For $k > 1$, the JPDF $f_k = f_k(t_1, \ldots, t_k)$ is given by*

$$f_k = \frac{1}{\pi^{k-1}\sqrt{2\pi}} \frac{1}{(t_1^2 + \cdots + t_k^2)} \frac{1}{\sqrt{t_1^2 + \cdots + t_{k-1}^2}} \cdots \frac{1}{\sqrt{t_1^2 + t_2^2}} \frac{\exp(-1/2t_1^2)}{|t_1|}.$$

3. Moments of T_k. In this section and the next, L_n continues to represent a triangular matrix of dimension n with $N(0,1)$ entries. As we remarked earlier, the exponential growth of $\kappa_n = \|L_n\|_2 \|L_n^{-1}\|_2$ is due to the second factor $\|L_n^{-1}\|_2$. Since the 2-norm of column $i+1$ of L_n^{-1} has the same distribution as T_{n-i}, we derive formulas for various moments of T_k with the intention of understanding the exponential growth of $\|L_n^{-1}\|_2$ with n.

In the lemma below, we consider the expected value $E(T_k^\xi)$ for both positive and negative values of ξ. By our notation, $T_1 = |t_1|$. The notation $d\Omega_k = dt_k \dots dt_1$ is used to reduce clutter in the proof. As usual, R^k denotes the real Euclidean space of dimension k.

The next lemma is stated as a recurrence to reflect the structure of its proof. Lemma 3.2 contains the same information in a simpler form.

LEMMA 3.1. *For any real $\xi < 1$, $E(T_k^\xi)$ is given by the following recurrence:*

$$E(T_1^\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\exp(-1/2x^2)}{|x|^{2-\xi}} dx, \tag{3.1}$$

$$E(T_k^\xi) = \frac{E(T_{k-1}^\xi)}{\pi} \int_{-\infty}^{\infty} \frac{dx}{(1+x^2)^{1-\xi/2}} \quad \textit{for } k > 1. \tag{3.2}$$

For $\xi \geq 1$ and $k \geq 1$, $E(T_k^\xi)$ is infinite.

Proof. To obtain (3.1), use $T_1 = |t_1|$ and the PDF of t_1 given by (2.2). It is easily seen that the integral is convergent if and only if $\xi < 1$.

Next, assume $k > 1$. By definition,

$$E(T_k^\xi) = \int_{R^k} T_k^\xi f_k d\Omega_k.$$

Using the recursive equation (2.3) for f_k, and writing T_k in terms of t_k and T_{k-1}, we get

$$\begin{aligned} E(T_k^\xi) &= \frac{1}{\pi} \int_{R^k} \frac{T_{k-1}}{T_k^{2-\xi}} f_{k-1} d\Omega_k \\ &= \frac{1}{\pi} \int_{R^{k-1}} \int_{-\infty}^{\infty} \frac{dt_k}{(t_k^2 + T_{k-1}^2)^{1-\xi/2}} T_{k-1} f_{k-1} d\Omega_{k-1}. \end{aligned} \tag{3.3}$$

By the substitution $t_k = xT_{k-1}$, the inner integral with respect to dt_k can be reduced to

$$T_{k-1}^{\xi-1} \int_{-\infty}^{\infty} \frac{dx}{(1+x^2)^{1-\xi/2}}.$$

Inserting this in the multiple integral (3.3) gives the recursive equation (3.2) for $E(T_k^\xi)$. It is easily seen that the integral in (3.2) is convergent if and only if $\xi < 1$. ∎

Define γ_ξ by

$$\gamma_\xi = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{dx}{(1+x^2)^{1-\xi/2}}. \tag{3.4}$$

Beginning with the substitution $x = \tan\theta$ in (3.4), it can be shown that $\gamma_\xi = \pi^{-1} B((1-\xi)/2, 1/2)$, where B is the beta function. The relevant expression for the beta function $B(x, y)$ is (6.2.1) in [1]. Also, if x is chosen from the standard Cauchy distribution, then $\gamma_\xi = E((1+x^2)^{\xi/2})$. We do not need γ_ξ in terms of the beta function, however; the integral expression (3.4) is more suitable for our purposes. Lemma 3.1 can be restated in a more convenient form using γ_ξ as follows.

LEMMA 3.2. *For $\xi < 1$, $E(T_k^\xi) = C_\xi \gamma_\xi^k$ for a finite positive constant C_ξ. Also, $\gamma_0 = 1$, $\gamma_\xi < 1$ for $\xi < 0$, and $\gamma_\xi > 1$ for $\xi > 0$.*

Proof. The expression for $E(T_k^\xi)$ is a restatement of Lemma 3.1. By elementary integration, $\gamma_0 = 1$, and by the form of the integral in (3.4), $\gamma_\xi < 1$ for $\xi < 0$ and $\gamma_\xi > 1$ for $\xi > 0$. ∎

Lemma 3.2 implies that the positive moments of T_k grow exponentially with k while the negative moments decrease exponentially with k.

Obtaining bounds for $P(T_k > M^k)$ and $P(T_k < m^k)$ is now a simple matter.

LEMMA 3.3. *For $k \geq 1$, $\xi > 0$, and $m > 0$,*

$$P(T_k < m^k) < C_{-\xi}(m/\gamma_{-\xi}^{-1/\xi})^{\xi k}.$$

Proof. Since $\xi > 0$, $P(T_k < m^k) = P(T_k^{-\xi} > m^{-\xi k})$. Use Lemma 3.2 with $\xi = -\xi$ to obtain an expression for $E(T_k^{-\xi})$ and apply Markov's inequality [2]. ∎

LEMMA 3.4. *For $k \geq 1$, $0 < \xi < 1$, and $M > 0$,*

$$P(T_k > M^k) < C_\xi(\gamma_\xi^{1/\xi}/M)^{\xi k}.$$

Proof. As in Lemma 3.3, $\xi > 0$ implies that $P(T_k > M^k) = P(T_k^\xi > M^{\xi k})$. Again, the proof can be completed by obtaining an expression for $E(T_k^\xi)$ using Lemma 3.2 followed by an application of Markov's inequality. ∎

4. Exponential growth of κ_n. We are now prepared to derive the first main result of the paper, namely, $\sqrt[n]{\kappa_n} \to 2$ almost surely as $n \to \infty$ for triangular matrices L_n with $N(0,1)$ entries. In what follows, *a.s.* means almost surely as $n \to \infty$. The definition of almost sure convergence for a sequence of random variables can be found in most textbooks on probability; for example, see [2]. Roughly, it means that the convergence holds for a set of sequences of measure 1.

LEMMA 4.1. $\|L_n\|_2^{1/n} \to 1$ *almost surely as $n \to \infty$.*

Proof. The proof is easy. We provide only an outline. The Frobenius norm of L_n, $\|L_n\|_F^2$, is a sum of $n(n+1)/2$ independent χ^2 variables of mean 1. By an argument exactly analogous to the proof of the strong law of large numbers with finite fourth moment assumption [2, p. 80],

$$\frac{\|L_n\|_F^2}{n(n+1)/2} \to 1 \ \ a.s.$$

The proof can be completed using the inequalities $n^{-1/2}\|L_n\|_F \leq \|L_n\|_2 \leq \|L_n\|_F$. Note that the suggested proof relies on the existence of the fourth moment of the χ^2 variables. ∎

The proof of Lemma 4.2 uses the first Borel–Cantelli lemma in a way that is typical of several proofs in probability. We use $\liminf_{n\to\infty} x_n$ and $\limsup_{n\to\infty} x_n$ for $\lim_{n\to\infty} \inf_{k\geq n} x_k$ and $\lim_{n\to\infty} \sup_{k\geq n} x_k$ in the following lemma and later.

LEMMA 4.2. *As $n \to \infty$, for any $0 < \xi < 1$,*

$$\gamma_{-\xi}^{-1/\xi} \leq \liminf_{n\to\infty} \sqrt[n]{\kappa_n} \leq \limsup_{n\to\infty} \sqrt[n]{\kappa_n} \leq \gamma_\xi^{1/\xi} \quad \textit{almost surely.}$$

Proof. By Lemma 4.1, it suffices to show that

$$\gamma_{-\xi}^{-1/\xi} \leq \liminf_{n\to\infty} \sqrt[n]{\|L_n^{-1}\|_2} \leq \limsup_{n\to\infty} \sqrt[n]{\|L_n^{-1}\|_2} \leq \gamma_\xi^{1/\xi} \ \ a.s.$$

We consider the lower bound first. The 2-norm of the first column of L_n^{-1}, which has the same distribution as T_n, is less than or equal to $\|L_n^{-1}\|_2$. Therefore, for $0 < \epsilon < 1$,

$$P\left(\sqrt[n]{\|L_n^{-1}\|_2} < \gamma_{-\xi}^{-1/\xi} - \epsilon\right) \leq P(T_n < (\gamma_{-\xi}^{-1/\xi} - \epsilon)^n).$$

Using Lemma 3.3 with $k = n$ and $m = \gamma_{-\xi}^{-1/\xi} - \epsilon$, we get

$$P\left(\sqrt[n]{\|L_n^{-1}\|_2} < \gamma_{-\xi}^{-1/\xi} - \epsilon\right) < C_{-\xi}\left(\frac{\gamma_{-\xi}^{-1/\xi} - \epsilon}{\gamma_{-\xi}^{-1/\xi}}\right)^{\xi n} = C_{-\xi} p_\epsilon^{\xi n},$$

where $p_\epsilon = \gamma_{-\xi}^{1/\xi}(\gamma_{-\xi}^{-1/\xi} - \epsilon) < 1$. Since $|p_\epsilon| < 1$, $\sum_{n=1}^{\infty} p_\epsilon^{\xi n}$ is finite. The first Borel–Cantelli lemma [2] can be applied to obtain

$$P\left(\sqrt[n]{\|L_n^{-1}\|_2} < \gamma_{-\xi}^{-1/\xi} - \epsilon \ \text{ infinitely often as } n \to \infty\right) = 0.$$

Taking the union of the sets in the above equation over all rational ϵ in $(0, 1)$ and considering the complement of that union, we obtain

$$P\left(\liminf_{n\to\infty} \sqrt[n]{\|L_n^{-1}\|_2} \geq \gamma_{-\xi}^{-1/\xi} \ \text{ as } n \to \infty\right) = 1.$$

In other words, $\gamma_{-\xi}^{-1/\xi} \leq \liminf_{n\to\infty} \sqrt[n]{\|L_n^{-1}\|_2}$ *a.s.*

The upper bound can be established similarly. At least one of the columns of L_n^{-1} must have 2-norm greater than or equal to $n^{-1/2}\|L_n^{-1}\|_2$. Since the 2-norm of column $k+1$ has the same distribution as T_{n-k},

$$P\left(\sqrt[n]{\|L_n^{-1}\|_2} > \gamma_\xi^{1/\xi} + \epsilon\right) \leq \sum_{k=1}^{n} P(T_k > n^{-1/2}(\gamma_\xi^{1/\xi} + \epsilon)^n).$$

Bounding each term in the summation using Lemma 3.4 gives

$$P\left(\sqrt[n]{\|L_n^{-1}\|_2} > \gamma_\xi^{1/\xi} + \epsilon\right) < C_\xi n^{\xi/2} \sum_{k=1}^{n} \left(\frac{\gamma_\xi^k}{(\gamma_\xi^{1/\xi} + \epsilon)^{\xi n}}\right).$$

Since $\gamma_\xi > 1$ by Lemma 3.2, the largest term in the summand occurs when $k = n$. Therefore,

$$P\left(\sqrt[n]{\|L_n^{-1}\|_2} > \gamma_\xi^{1/\xi} + \epsilon\right) < C_\xi n^{1+\xi/2}\left(\frac{\gamma_\xi^{1/\xi}}{\gamma_\xi^{1/\xi} + \epsilon}\right)^{\xi n}.$$

From this point, the proof can be completed in the same manner as the proof of the lower bound. □

THEOREM 4.3. *For random triangular matrices with $N(0, 1)$ entries, as $n \to \infty$,*

$$\sqrt[n]{\kappa_n} \to 2 \quad \textit{almost surely.}$$

Proof. By an inequality sometimes called Lyapunov's [13, p. 144], [2],

$$\gamma_\beta^{1/\beta} < \gamma_\alpha^{1/\alpha}$$

for any real $\beta < \alpha$. Thus, the bounding intervals $[\gamma_{-\xi}^{-1/\xi}, \gamma_\xi^{1/\xi}]$ in Lemma 4.2 shrink as ξ decreases from 1 to 0. A classical theorem [13, p. 139] says that these intervals actually shrink to the following point:

$$\lim_{\xi\to 0}\gamma_\xi^{1/\xi} = \lim_{\xi\to 0}\left(\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{(1+x^2)^{1-\xi/2}}dx\right)^{1/\xi}$$
$$= \exp\left(\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\log(1+x^2)}{1+x^2}dx\right).$$

The exact value of the limit can be evaluated to 2 using the substitution $x = \tan\theta$ followed by complex integration [3, p. 121]. Thus, $\sqrt[n]{\kappa_n} \to 2$ *a.s.* □

Theorem 4.3 holds in exactly the same form if the nonzero entries of L_n are independent $N(0,\sigma^2)$ variables rather than $N(0,1)$ variables, since the condition number is invariant under scaling.

Our approach to Theorem 4.3 began by showing that $E(T_k^\xi) = C_\xi\gamma_\xi^k$ for both positive and negative ξ. Once these expressions for the moments of T_k were obtained, our arguments did not depend on how the recurrence was computed. The following note summarizes the asymptotic information about a recurrence that can be obtained from a knowledge of its moments.

Note. Let $t_1, t_2, \ldots$ be a sequence of random variables. If $E(|t_n|^\xi)$ grows exponentially with n at the rate ν_ξ^n for $\xi > 0$, then $\limsup_{n\to\infty}\sqrt[n]{|t_n|} \le \nu_\xi^{1/\xi}$ almost surely. Similarly, if $E(|t_n|^\xi)$ decreases exponentially with n at the rate ν_ξ^n as $n \to \infty$ for $\xi < 0$, then $\nu_\xi^{1/\xi} \le \liminf_{n\to\infty}\sqrt[n]{|t_n|}$ almost surely. Thus, knowledge of any positive moment of t_n yields an upper bound on $\sqrt[n]{|t_n|}$ as $n \to \infty$, while knowledge of any negative moment yields a lower bound.

5. Unit triangular matrices. So far, we have considered triangular matrices whose nonzero entries are independent, real $N(0,1)$ variables. In this section and in section 7, we establish the exponential growth of the condition number for other kinds of random triangular matrices with normally distributed entries. The key steps in the sequence of lemmas leading to the analogues of Theorem 4.3 are stated but not proved. The same techniques used in sections 2, 3, and 4 work here, too.

Let L_n be a unit lower triangular matrix of dimension n with $N(0,\sigma^2)$ subdiagonal entries. Let $s_1, \ldots, s_k$ be the first k entries from the diagonal downwards of any column of L_n^{-1}. The entries s_i satisfy the recurrence

$$\begin{aligned} s_1 &= 1,\\ s_2 &= \alpha_{21}s_1,\\ s_3 &= \alpha_{31}s_1 + \alpha_{32}s_2,\\ &\vdots\\ s_k &= \alpha_{k1}s_1 + \cdots + \alpha_{k,k-1}s_{k-1}, \end{aligned} \tag{5.1}$$

where α_{ij}, $i > j$, are $N(0,\sigma^2)$ variables. The notation $S_k = \sqrt{s_1^2 + s_2^2 + \cdots + s_k^2}$ is used below.

PROPOSITION 5.1. *The JPDF of $s_1, \ldots, s_k$, $g_k(s_1, \ldots, s_k)$, is given by the recurrence*

$$g_2 = \frac{1}{\sqrt{2\pi}\sigma}\exp(-s_2^2/2\sigma^2),$$
$$g_k = \frac{1}{\sqrt{2\pi}\sigma}\frac{\exp(-s_k^2/2\sigma^2 S_{k-1}^2)}{S_{k-1}}g_{k-1} \quad \textit{for } k > 2,$$

and the fact that $s_1 = 1$ identically.

LEMMA 5.2. *For any real* ξ, $E(S_k^\xi) = \lambda_\xi^{k-1}$, *where*

$$\lambda_\xi = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (1+x^2)^{\xi/2} \exp(-x^2/2\sigma^2) dx.$$

The note at the end of section 4 provides part of the link from Lemma 5.2 to the following theorem about κ_n.

THEOREM 5.3. *For random unit triangular matrices with* $N(0,\sigma^2)$ *entries, as* $n \to \infty$,

$$\sqrt[n]{\kappa_n} \to \exp\left(\frac{1}{2\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \log(1+x^2) e^{-x^2/2\sigma^2} dx\right) \quad \textit{almost surely.}$$

If this limit is denoted by $C(\sigma)$, *then*

$$C(\sigma) - 1 \sim \sigma^2/2 \quad \textit{as } \sigma \to 0,$$
$$C(\sigma) \sim K\sigma \quad \textit{as } \sigma \to \infty,$$

where $K = \sqrt{\exp(-\gamma)/2} = 0.5298\ldots$, *with* γ *being the Euler constant.*

Proof. The constant K is given by

$$K = \exp\left(\sqrt{\frac{2}{\pi}} \int_0^{\infty} \log x \exp(-x^2/2) dx\right).$$

To evaluate K, we used integral 4.333 of [8]. □

In contrast to the situation in Theorem 4.3, the constant that $\sqrt[n]{\kappa_n}$ converges to in Theorem 5.3 depends on σ. This is because changing σ scales only the subdiagonal entries of the unit triangular matrix L_n while leaving the diagonal entries fixed at one. For $\sigma = 1$, the case discussed in the Introduction, numerical integration shows the constant to be $1.305683410\ldots$.

6. A comment on the stability of Gaussian elimination. The conditioning of random unit triangular matrices has a connection with the phenomenon of numerical stability of Gaussian elimination. We pause briefly to explain this connection.

For decades, the standard algorithm for solving general systems of linear equations $Ax = b$ has been Gaussian elimination (with "partial" or row pivoting). This algorithm generates an "LU factorization" $PA = LU$, where P is a permutation matrix, L is unit lower triangular with subdiagonal entries less than or equal to one in absolute value, and U is upper triangular.

In the mid-1940s it was predicted by Hotelling [14] and Goldstine and von Neumann [9] that rounding errors must accumulate exponentially in elimination algorithms of this kind, causing instability for all but small dimensions. In the 1950s, Wilkinson developed a beautiful theory based on backward error analysis that, while it explained a great deal about Gaussian elimination, confirmed that for certain matrices, exponential instability does indeed occur [19]. He showed that amplification of rounding errors by factors on the order of $\|L^{-1}\|$ may take place, and that for certain matrices, $\|L^{-1}\|$ is of order 2^n. Thus, for certain matrices, rounding errors are amplified by $O(2^n)$, causing a catastrophic loss of n bits of precision.

Despite these facts, the experience of 50 years of computing has established that from a practical point of view, Hotelling and von Neumann were wrong: Gaussian elimination is overwhelmingly stable. In fact, it is not clear that a single matrix

problem has ever led to an instability in this algorithm, except for the ones produced by numerical analysts with that end in mind, although Foster [7] and Wright [20] have devised problems leading to instability that plausibly "might have arisen" in applications. The reason appears to be statistical: the matrices A for which $\|L^{-1}\|$ is large occupy an exponentially small proportion of the space of all matrices, so small that such matrices "never" arise in practice. Experimental evidence of this phenomenon is presented in [18].

This raises the question, why are matrices A for which $\|L^{-1}\|$ is large so rare? It is here that the behavior of random unit triangular matrices is relevant. A natural hypothesis would be that the matrices L generated by Gaussian elimination are, to a reasonable approximation, random unit triangular matrices with off-diagonal entries of a size dependent on the dimension n. If such matrices could be shown to be almost always well conditioned, then the stability of Gaussian elimination would be explained.

We have just shown, however, that unit triangular matrices are exponentially ill conditioned. Thus, this attempted explanation of the stability of Gaussian elimination fails, and indeed, the same argument suggests that Gaussian elimination should be unstable in practice as well as in the worst case. The resolution of this apparent paradox is that *the matrices L produced by Gaussian elimination are far from random.* The signs of the entries of these matrices are correlated in special ways that have the effect of keeping $\|L^{-1}\|$ almost always very small. For example, it is reported in [18] that a certain random matrix A with $n = 256$ led to $\|L^{-1}\| = 33.2$, whereas if $\tilde{L}$ was taken to be the same matrix but with the signs of its subdiagonal entries randomized, the result became $\|\tilde{L}^{-1}\| = 2.7 \times 10^8$. In fact, even unpivoted Gaussian elimination does not produce triangular matrices as severely ill conditioned as random triangular matrices [22].

From a comparison of Theorem 5.3 with half a century of the history of Gaussian elimination, then one may conclude that unit triangular factors of random dense matrices are very different from random unit triangular matrices. An explanation of this difference is offered in [17] along the following lines. If A is random, then its successive column spaces are randomly oriented in n-space in the sense that the first column of A is oriented in a random direction, the span of the first two columns is a random two-dimensional space, and so on. Since the span of the first k columns of L is the same as the span of the first k columns of PA, where P is the row permutation matrix produced by partial pivoting, the same holds approximately for the successive column spaces of L. That condition, in turn, implies that large values $\|L^{-1}\|$ can arise only exponentially rarely.

7. Complex matrices. We now consider matrices with complex entries. Let L_n be a lower triangular matrix with $\tilde{N}(0,1)$ entries. The complex distribution $\tilde{N}(0,1)$ was defined in the Introduction. Let $t_1, \ldots, t_k$ denote the first k entries from the diagonal downwards of any column of L_n^{-1}. The quantities t_k satisfy (2.1), but the α_{ij} are now independent $\tilde{N}(0,1)$ variables. Let $r_k = |t_k|^2$, and denote $r_1 + \cdots + r_k$ by R_k.

PROPOSITION 7.1. *The JPDF of $r_1, \ldots, r_k$, $h_k(r_1, \ldots, r_k)$, is given by the recurrence*

$$h_1 = \frac{\exp(-1/r_1)}{r_1^2}, \tag{7.1}$$

$$h_k = \frac{R_{k-1}}{R_k^2} h_{k-1} \quad \textit{for } k > 1, \tag{7.2}$$

for $r_i \geq 0$, $1 \leq i \leq k$.

Proof. We sketch only the details that do not arise in the proof of Proposition 2.1. If x and y are independent $N(0, \sigma^2)$ variables, $x = \sqrt{r}\cos(\theta)$ and $y = \sqrt{r}\sin(\theta)$, then r and θ are independent. Moreover, the distribution of r is exponential with the PDF

$$(1/2\sigma^2)\exp(-r/2\sigma^2) \tag{7.3}$$

for $r > 0$.

Consider the sum $\tau_k = \alpha_{k1}t_1 + \cdots + \alpha_{k,k-1}t_{k-1}$ with α_{ki} taken as independent $\tilde{N}(0,1)$ variables. For fixed $t_1, \ldots, t_{k-1}$, $\mathrm{Re}(\tau_k)$ and $\mathrm{Im}(\tau_k)$ are independent. To see their independence, we write out the equations for $\mathrm{Re}(\tau_k)$ and $\mathrm{Im}(\tau_k)$ as follows:

$$\mathrm{Re}(\tau_k) = \sum_{i=1}^{k-1} \mathrm{Re}(\alpha_{ki})\mathrm{Re}(t_i) - \mathrm{Im}(\alpha_{ki})\mathrm{Im}(t_i),$$

$$\mathrm{Im}(\tau_k) = \sum_{i=1}^{k-1} \mathrm{Re}(\alpha_{ki})\mathrm{Im}(t_i) + \mathrm{Im}(\alpha_{ki})\mathrm{Re}(t_i).$$

The linear combinations of $\mathrm{Re}(\alpha_{ki})$ and $\mathrm{Im}(\alpha_{ki})$ in these two equations can be realized by taking inner products with the two vectors

$$v = [\mathrm{Re}(t_1), \ldots, \mathrm{Re}(t_{k-1}), -\mathrm{Im}(t_1), \ldots, -\mathrm{Im}(t_{k-1})],$$

$$w = [\mathrm{Im}(t_1), \ldots, \mathrm{Im}(t_{k-1}), +\mathrm{Re}(t_1), \ldots, +\mathrm{Re}(t_{k-1})].$$

The independence of $\mathrm{Re}(\tau_k)$ and $\mathrm{Im}(\tau_k)$ is a consequence of the orthogonality of v and w, i.e., $(v, w) = vw' = 0$, and the invariance of the JPDF of independent, identically distributed normal variables under orthogonal transformation [15].

Thus, for fixed $t_1, \ldots, t_{k-1}$, the real and imaginary parts of τ_k are independent normal variables of mean 0 and variance $R_{k-1}/2$. By (7.3), the PDFs of $x = |\tau_k|^2$ and $y = |\alpha_{kk}|^2$ are given by

$$\frac{1}{R_{k-1}} \exp(-x/R_{k-1}), \quad \exp(-y)$$

for positive x, y. The expression (7.2) for h_k can now be obtained using $r_k = |\tau_k|^2/|\alpha_{kk}|^2$. □

LEMMA 7.2. *For any* $\xi < 1$, $E(R_k^\xi) = C\mu_\xi^{k-1}$, *where*

$$C = \int_0^\infty \frac{\exp(-1/r_1)}{r_1^{2-\xi}} dr_1, \quad \mu_\xi = \int_0^\infty \frac{dx}{(1+x)^{2-\xi}}.$$

The constant μ_ξ in Lemma 7.2 can be reduced to $(1-\xi)^{-1}$ for $\xi < 1$. However, as with γ_ξ in section 3, the integral expression for μ_ξ is more suitable for our purposes. As before, the note at the end of section 4 is an essential part of the link from the previous lemma to the following theorem about κ_n.

THEOREM 7.3. *For random triangular matrices with complex $\tilde{N}(0,1)$ entries, as $n \to \infty$,*

$$\sqrt[n]{\kappa_n} \to \exp\left(\frac{1}{2}\int_0^\infty \frac{\log(1+x)}{(1+x)^2}dx\right) = e^{1/2} \quad \textit{almost surely.}$$

Theorem 7.3 holds unchanged if the entries are $\tilde{N}(0,\sigma^2)$ variables. As with Theorem 4.3, this is because the condition number is invariant under scaling.

Now, let L_n be a unit lower triangular matrix of dimension n with $\tilde{N}(0,\sigma^2)$ subdiagonal entries. We state only the final theorem about κ_n.

THEOREM 7.4. *For random unit triangular matrices with complex $\tilde{N}(0,\sigma^2)$ entries, as $n \to \infty$,*

$$\begin{aligned}\sqrt[n]{\kappa_n} &\to \exp\left(\frac{1}{4}\int_0^\infty \log(1+\sigma^2 x/2)e^{-x/2}dx\right)\\ &= \exp(-\exp(\sigma^{-2})\mathrm{Ei}(-\sigma^{-2})/2) \quad \textit{almost surely,}\end{aligned}$$

where Ei *is the exponential integral. If this limit is denoted by $C(\sigma)$, then*

$$\begin{aligned}C(\sigma) - 1 &\sim \sigma^2/2 \quad \textit{as } \sigma \to 0,\\ C(\sigma) &\sim K\sigma \quad \textit{as } \sigma \to \infty,\end{aligned}$$

where $K = \exp(-\gamma/2) = 0.7493\ldots$, with γ being the Euler constant.

Proof. To obtain K, we evaluated

$$K = \exp\left(\frac{1}{4}\int_0^\infty \log(x/2)\exp(-x/2)dx\right)$$

using the Laplace transform of $\log(x)$ given by integral 4.331.1 of [8]. The explicit formula involving $\mathrm{Ei}(\sigma^{-2})$ was obtained using integral 4.337.2 of [8]. ∎

For $\sigma^2 = 1$, $\sqrt[n]{\kappa_n}$ converges to $1.347395784\ldots$.

8. Matrices with entries from stable distributions. The techniques used to deduce Theorem 4.3 require that we first derive the joint density function of the t_k, defined by recurrence (2.1), as was done in Proposition 2.1. That proposition made use of the fact that when the α_{ki} are independent and normally distributed, and the t_i are fixed, the sum

$$\sum_{i=1}^{k-1} \alpha_{ki} t_i$$

is also normally distributed. This property of the normal distribution holds for any stable distribution.

A distribution is said to be stable if for X_i chosen independently from that distribution,

$$\sum_{i=1}^{n} X_i$$

has the same distribution as $c_n X + d_n$, where X has the same distribution as X_i and $c_n > 0$ and d_n are constants [6, p. 170]. If $d_n = 0$, the distribution is said to be

strictly stable. As usual, the distribution is symmetric if X has the same distribution as $-X$. A symmetric, strictly stable distribution has exponent a if $c_n = n^{1/a}$. A standard result of probability theory says that any stable distribution has an exponent $0 < a \leq 2$. The normal distribution is stable with exponent $a = 2$ [6].

The techniques used for triangular matrices with normal entries work more generally when the entries are drawn from a symmetric, strictly stable distribution. Let L_n be a unit lower triangular matrix with entries chosen from a symmetric, strictly stable distribution. Denote the PDF of that stable distribution by $\phi(x)$. The recurrence for the entries s_i of the inverse L_n^{-1} is again given by (5.1), but α_{ki}, $k > i$, are now independent random variables with the density function $\phi(x)$.

Our program for deriving the constants that $\sqrt[n]{\kappa_n}$ converge to as $n \to \infty$ began with Lemma 4.1 in all the previous examples. One of referees pointed out to us that a new proof is needed for that lemma in the present context since a stable distribution of index $a < 2$ does not have the ath or higher moments.

LEMMA 8.1. *For $a < 2$, $\|L_n\|_2^{1/n} \to 1$ almost surely as $n \to \infty$.*

Proof. Define $\|L_n\|_\alpha = (\sum_{i,j} |l_{ij}|^\alpha)^{1/\alpha}$ for some $0 < \alpha < a/4$. Then the inequality

$$n^{(1/2-2/\alpha)} \|L_n\|_\alpha \leq \|L_n\|_2 \leq n \|L_n\|_\alpha$$

and the existence of the fourth moment of $|l_{ij}|^\alpha$ make possible a proof analogous to what was outlined for Lemma 4.1. □

The proposition, the lemma, and the theorem below are analogues of Proposition 5.1, Lemma 5.2, and Theorem 5.3, respectively. If the exponent of the stable distribution is a, denote $(|s_1|^a + \cdots + |s_k|^a)^{1/a}$ by S_k.

PROPOSITION 8.2. *If $\phi(x)$ is the density function of a symmetric, strictly stable distribution with exponent a, the JPDF of $s_1, \ldots, s_k$, $g_k(s_1, \ldots, s_k)$, is given by the recurrence*

$$g_2 = \phi(s_2),$$

$$g_k = \frac{\phi(s_k/S_{k-1})}{S_{k-1}} g_{k-1} \quad \text{for } k > 2,$$

and the fact that $s_1 = 1$ identically.

Proof. The proof is very similar to the proof of Proposition 2.1. We note that if α_{ki}, $k > i$, are independent random variables with the PDF $\phi(x)$, and the s_i are fixed, then the sum

$$\alpha_{k1} s_1 + \cdots + \alpha_{k,k-1} s_{k-1}$$

has the PDF $\phi(x/S_{k-1})/S_{k-1}$ [6, p. 171]. □

LEMMA 8.3. *For any real ξ, $E(S_k^\xi) = \lambda_\xi^{k-1}$, where*

$$\lambda_\xi = \int_{-\infty}^{+\infty} (1 + |x|^a)^{\xi/a} \phi(x) dx,$$

with $\lambda_\xi = \infty$ for $\xi \geq a$.

THEOREM 8.4. *For random unit triangular matrices with entries from a symmetric, strictly stable distribution with density function $\phi(x)$ and exponent a, as $n \to \infty$,*

$$\sqrt[n]{\kappa_n} \to \exp\left(\frac{1}{a} \int_{-\infty}^{\infty} \log(1 + |x|^a) \phi(x) dx\right) \quad \text{almost surely.}$$

Theorem 5.3 is a special case of Theorem 8.4 when $\phi(x)$ is the density function for the symmetric, strictly stable distribution $N(0, \sigma^2)$. Another notable symmetric, strictly stable distribution is the Cauchy distribution with the density function

$$\phi(x) = \frac{1}{\pi} \frac{1}{1 + x^2}.$$

The exponent a for the Cauchy distribution is 1 [6]. Using Theorem 8.4 we obtain the following.

THEOREM 8.5. *For random unit triangular matrices with entries from the standard Cauchy distribution, as* $n \to \infty$,

$$\sqrt[n]{\kappa_n} \to \exp\left(\frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{\log(1 + |x|)}{1 + x^2} dx\right) \quad \textit{almost surely.}$$

Numerical integration shows the constant to be $2.533737279\ldots$.

A similar generalization can be made for triangular matrices without a unit diagonal. However, the analogue of Theorem 8.4 for such matrices involves not $\phi(x)$ but the density function $\psi(x)$ of the quotient $x = y/z$ obtained by taking y, z as independent variables with the PDF ϕ. The distribution ψ can be difficult to compute and work with. We state only the final theorem about κ_n for triangular matrices with entries drawn from a symmetric strictly stable distribution.

THEOREM 8.6. *For random triangular matrices with entries from a symmetric, strictly stable distribution with density function* $\phi(x)$ *and exponent* a, *as* $n \to \infty$,

$$\sqrt[n]{\kappa_n} \to \exp\left(\frac{1}{a} \int_{-\infty}^{\infty} \log(1 + |x|^a) \psi(x) dx\right) \quad \textit{almost surely,}$$

where $\psi(x)$ *is the density function of the quotient of two independent variables with the density function* $\phi(x)$.

Theorem 4.3 is a special case of Theorem 8.6 when $\phi(x)$ is the density function of the distribution $N(0, \sigma^2)$. The $\psi(x)$ corresponding to $N(0, \sigma^2)$ is the standard Cauchy distribution. To apply Theorem 8.6 for the Cauchy distribution, we note that

$$\psi(x) = \frac{2}{\pi^2} \frac{\log|x|}{x^2 - 1}$$

is the density function of the quotient if the numerator and the denominator are independent Cauchy variables [11]. Therefore, Theorem 8.6 implies the following.

THEOREM 8.7. *For random triangular matrices with entries from the standard Cauchy distribution, as* $n \to \infty$,

$$\sqrt[n]{\kappa_n} \to \exp\left(\frac{2}{\pi^2} \int_{-\infty}^{\infty} \log(1 + |x|) \frac{\log|x|}{x^2 - 1} dx\right) \quad \textit{almost surely.}$$

The constant of convergence in Theorem 8.7 is $3.063094192\ldots$.

9. Summary. Below is a summary of the exponential growth factors $\lim_{n\to\infty} \sqrt[n]{\kappa_n}$ that we have established for triangular matrices with normal entries:

Real triangular	2	Theorem 4.3
Real unit triangular, $\sigma^2 = 1$	$1.305683410\ldots$	Theorem 5.3
Complex triangular	$e^{1/2} = 1.647\ldots$	Theorem 7.3
Complex unit triangular, $\sigma^2 = 1$	$1.347395784\ldots$	Theorem 7.4

The theorems about unit triangular matrices with normally distributed, real or complex entries apply for any variance σ^2, not just $\sigma^2 = 1$. Constants of convergence for any symmetric, strictly stable distribution were derived in Theorems 8.4 and 8.6. Those two theorems were specialized to the Cauchy distribution in Theorems 8.5 and 8.7.

Similar results seem to hold more generally, i.e., even when the entries of the random triangular matrix are not from a stable distribution. Moreover, the complete knowledge of moments achieved in Lemma 3.2 and its analogues might be enough to prove stronger limit theorems than Theorem 4.3 and its analogues. We will present limit theorems and results about other kinds of random triangular matrices in a later publication. We will also discuss the connection between random recurrences and products of random matrices, and the pseudospectra of infinite random triangular matrices.

For the random recurrences we have considered here, every new term is a random sum of all the previous terms in the sequence. The exponential increase of successive terms with probability 1 holds even for some random recurrences that generate a new term as a random sum of a fixed number of previous terms. For example, if we define random Fibonacci sequences by $t_1 = t_2 = 1$, and for $n > 2$, $t_n = \pm t_{n-1} \pm t_{n-2}$, where each $\pm$ sign is independent and either $+$ or $-$ with probability $1/2$, then $\sqrt[n]{|t_n|} \to 1.13198824\ldots$ almost surely [21]. Thus, the condition number increases exponentially even for some random triangular matrices that are banded.

We close with two figures that illustrate the first main result of this paper, namely, for random triangular matrices with $N(0,1)$ entries, $\sqrt[n]{\kappa_n} \to 2$ almost surely as $n \to \infty$ (Theorem 4.3). Figure 9.1 plots the results of a single run of the random recurrence (2.1) to $100,000$ steps, confirming the constant 2 to about two digits. The expense involved in implementing the full recurrence (2.1) for so many steps would be prohibitive. However, since t_k grows at the rate 2^k, we need include only a fixed number of terms in (2.1) to compute t_k to machine precision. For the figure, we used 200 terms, although half as many would have been sufficient. Careful scaling was necessary to avoid overflow while computing this figure.

Figure 9.2 plots the condition number of a single random triangular matrix for each dimension from 1 to 200. The exponential trend at the rate 2^n is clear, but as in Figure 1.1, the convergence as $n \to \infty$ is slow.

Acknowledgments. We thank D. Coppersmith, P. Diaconis, H. Kesten, A. Odlyzko, J. Sethna, H. Wilf, and the referees for helpful discussions and comments. We are especially grateful to Prof. Diaconis for introducing us to stable distributions and to one of the referees for a very careful reading of this paper.

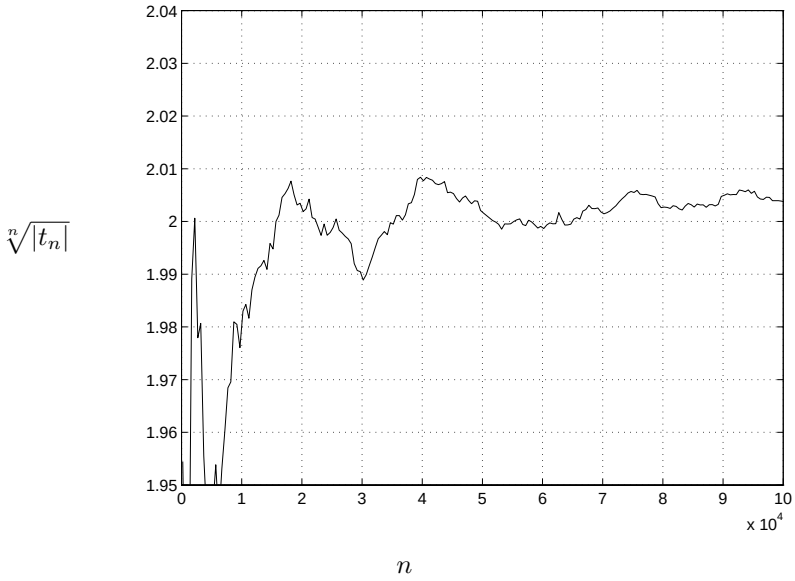


FIG. 9.1. *Illustration of Theorem* 4.3. *After* 100,000 *steps of the random recurrence* (2.1), $\sqrt[n]{|t_n|}$ *has settled to within* 1% *of its limiting value* 2. *The implementation is explained in the text.*

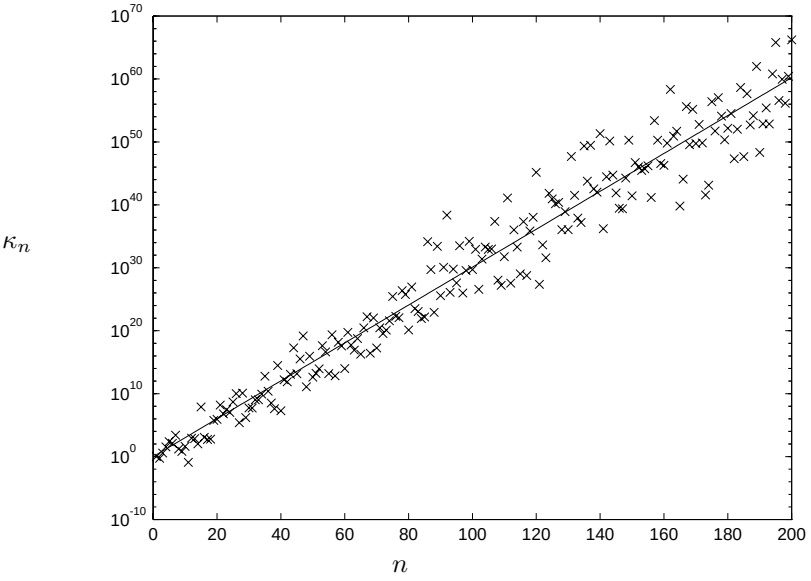


FIG. 9.2. *Another illustration of Theorem* 4.3. *Each cross is obtained by computing the condition number* κ_n *for one random triangular matrix of dimension* n *with* $N(0,1)$ *entries. The solid line represents* 2^n.

REFERENCES

[1] M. ABRAMOWITZ AND I.A. STEGUN, EDS., *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] P. BILLINGSLEY, *Probability and Measure*, 2nd ed., John Wiley, New York, 1986.

[3] J. B. Conway, *Functions of One Complex Variable*, Springer, New York, 1995.
[4] A. Edelman, *Eigenvalues and Condition Numbers of Random Matrices*, Ph.D. dissertation and Numerical Analysis Report 89-7, Massachusetts Institute of Technology, Cambridge, MA, 1989.
[5] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 543–560.
[6] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd ed., John Wiley, New York, 1971.
[7] L. V. Foster, *Gaussian elimination with partial pivoting can fail in practice*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1354–1362.
[8] I. S. Gradshteyn and I. W. Ryzhik, *Table of Integrals, Series, and Products*, 4th ed., Academic Press, New York, 1965.
[9] H. H. Goldstine and J. von Neumann, *Numerical inverting of matrices of high order*, Amer. Math. Soc. Bull., 53 (1947), pp. 1021–1099.
[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[11] A. K. Gupta and Z. Govindarajulu, *Distribution of the quotient of two independent Hotelling T2-variates*, Comm. Statist., 4 (1975), pp. 449–453.
[12] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[13] G. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, 2nd ed., Cambridge University Press, Cambridge, UK, 1988.
[14] H. Hotelling, *Some new methods in matrix calculation*, Ann. Math. Statist., 14 (1943), pp. 1–34.
[15] M. L. Mehta, *Random Matrices and the Statistical Theory of Energy Levels*, Academic Press, New York, 1967.
[16] J. W. Silverstein, *The smallest eigenvalue of a large-dimensional Wishart matrix*, Ann. Probab., 13 (1985), pp. 1364–1368.
[17] L. N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
[18] L. N. Trefethen and R. S. Schreiber, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 335–360.
[19] J. H. Wilkinson, *Error Analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.
[20] S. J. Wright, *A collection of problems for which Gaussian elimination with partial pivoting is unstable*, SIAM J. Sci. Comput., 14 (1993), pp. 231–238.
[21] D. Viswanath, *Random Fibonacci sequences and the number* 1.13198824 . . . , Math. Comp., submitted.
[22] M. C. Yeung and T. F. Chan, *Probabilistic analysis of Gaussian elimination without pivoting*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 499–517.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 583–612, July 1998

001

INERTIAS OF BLOCK BAND MATRIX COMPLETIONS*

NIR COHEN† AND JEROME DANCIS‡

Abstract. This paper classifies the ranks and inertias of hermitian completion for the partially specified 3×3 block band hermitian matrix (also known as a "bordered matrix")

$$P = \begin{pmatrix} A & B & ? \\ B^* & C & D \\ ? & D^* & E \end{pmatrix}.$$

The full set of completion inertias is described in terms of seven linear inequalities involving inertias and ranks of specified submatrices. The minimal completion rank for P is computed.

We study the completion inertias of partially specified hermitian block band matrices, using a block generalization of the Dym–Gohberg algorithm. At each inductive step, we use our classification of the possible inertias for hermitian completions of bordered matrices. We show that when all the maximal specified submatrices are invertible, any inertia consistent with Poincaré's inequalities is obtainable. These results generalize the nonblock band results of Dancis [*SIAM J. Matrix Anal. Appl.*, 14 (1993), pp. 813–829].

All our results remain valid for real symmetric completions.

Key words. matrices, hermitian, rank, inertia, completion, minimal rank

AMS subject classification. 15A57

PII. S0895479895296471

1. Introduction.

We address the following completion problem: given a partially specified hermitian matrix P, characterize all the possible inertias $\text{In } H = (p, n, d)$ of the various hermitian completions H of P. We call this set the "inertial set" or "inertial polygon" of P.

The issue of classifying positive definite and semidefinite completions of partial matrices is relevant to various applications involving interpolation and has been studied thoroughly, e.g., [AHMR], [D5], [GJSW]. Invertible completions have been studied in [DG] and [EGL2], for band patterns, in [L] for general patterns, and are associated with maximum entropy and statistics. For other results concerning ranks and general inertias, see [D1], [D2], [D3], [D4], [D5], [D6], [EL], [G], [H], [JR1], [CG3], [BJL], [CG1], [CG2], [D7], [HO], and [I].

Following some preliminary material (sections 2–4), we present in sections 5 and 6 several contributions to the inertia classification problem.

$\text{In}(H) = (\pi(H), \nu(H), \delta(H))$ will denote the inertia, that is, the number of positive, negative, and zero eigenvalues of a hermitian matrix H. They are also called the *positivity, negativity*, and *nullity* of H.

The main result is the following.

*Received by the editors December 21, 1995; accepted for publication (in revised form) by V. Mehrmann April 28, 1997; published electronically March 18, 1998. An earlier version of this paper was Technical Report 795, Department of Electrical Engineering, Technion - Israel Institute of Technology, 1991. Some of the results were announced at the Seventh Haifa Matrix Theory Conference, Haifa, Israel, 1991.

http://www.siam.org/journals/simax/19-3/29647.html

†DMA/IMECC, Unicamp, CP 6065, CEP 13083-970, Campinas SP, Brazil (nir@ ime.unicamp.br).

‡Department of Mathematics, University of Maryland, College Park, MD 20742-4015 (jdancis@math.umd.edu).

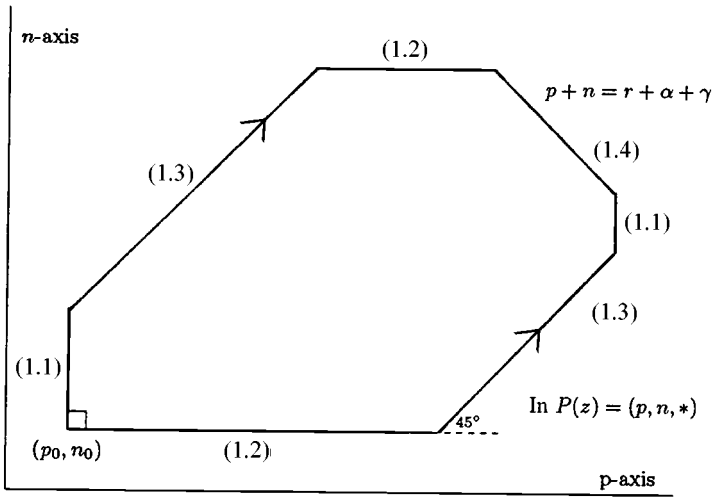


FIG. 1.1. *Graph of the inertial polygon for the bordered matrix.*

THEOREM 1.1. *Given the block "bordered" matrix*

$$P(Z) = \text{In} \begin{pmatrix} A & B & Z \\ B^* & C & D \\ Z^* & D^* & E \end{pmatrix},$$

where A, B, C, *and* E *are hermitian (real or complex) matrices of sizes* $\alpha \times \alpha$, $\beta \times \beta$, *and* $\gamma \times \gamma$, *respectively, and* B *and* D *are matrices of sizes* $\alpha \times \beta$ *and* $\beta \times \gamma$, *respectively, set*

$$(\pi, \nu, \delta) = \text{In} \begin{pmatrix} A & B \\ B^* & C \end{pmatrix} \quad \textit{and} \quad (\pi', \nu', \delta') = \text{In} \begin{pmatrix} C & D \\ D^* & E \end{pmatrix},$$

$$r = \text{rank}(B^* \ C \ D),$$

$$\Delta' = r - \text{rank}(B^* \ C) \quad \textit{and} \quad \Delta'' = r - \text{rank}(C \ D),$$

$$p_0 = \max\{\pi + \Delta', \pi' + \Delta''\} \quad \textit{and} \quad n_0 = \max\{\nu + \Delta', \nu' + \Delta''\}.$$

For given integers n *and* p, *there exist an* $\alpha \times \gamma$ *(real or complex, respectively) matrix* Z *such that*

$$\text{In } P(Z) = (p, n, \alpha + \beta + \gamma - n - p)$$

if and only if

$$p_0 \le p \le \min\{\alpha + \pi', \gamma + \pi, \}, \tag{1.1}$$

$$n_0 \le n \le \min\{\alpha + \nu', \gamma + \nu\}, \tag{1.2}$$

$$r - \nu - \nu' \le p - n \le \pi + \pi' - r, \tag{1.3}$$

$$p + n \le \alpha + \gamma + r. \tag{1.4}$$

The block partition is not required to be uniform; rectangular (nonsquare) blocks are permitted. For this partial matrix $P(Z)$, Theorem 1.1 shows that the inertial set is a (possibly degenerate) convex seven-sided lattice polygon as depicted in Fig. 1.1.

The proof of Theorem 1.1 is presented in section 5, along with a variety of corollaries including a small application to the algebraic matrix Riccati equation.

Cain and Sa established the 2×2 case (i.e., $\beta = 0$) in [CS]. The result was generalized to an arbitrary number of diagonal blocks by Cain in [C], with further results by Dancis in [D1]. The 2×2 case with one given diagonal block was proven as Theorem 1 of [S] and as Theorem 1.2 of [D1]. These cases are reviewed in detail in section 4, and are later used as milestones in computing the inertial polygon of Theorem 1.1.

The possible inertias for a bordered matrix missing a single (scalar) entry were cataloged by the second author in [D6], mostly using his extended Poincaré's inequalities (3.3). In [JR] the lower bounds in (5.5) and (5.6) were determined for the case when the given principal blocks are invertible (i.e., $\alpha+\beta = \pi+\nu$ and $\beta+\gamma = \pi'+\nu'$). Their result extends to the case of "chordal patterns."

In computing the inertial set for Theorem 1.1, we combine four simple elements: (i) Schur complements, (ii) Poincaré and Extended Poincaré Inequalities (3.3) as necessary conditions on the inertia, (iii) the technique of "restricted congruence" (presented in section 2.4), including a new formula (2.7) for simplifying a partial hermitian matrix, and (iv) elimination of variables in systems of linear inequalities (see sections 2–3 for details). These techniques enable us to reduce Theorem 1.1 to a combination of the simpler cases presented in section 4. These four elements of the proof, without (2.7), are commonly used in the matrix literature, and in particular in the completion literature cited above.

Staircase hermitian matrices are the mild generalization of block band matrices described as *generalized block band matrices* in the appendix of [JR2]; they look like a double staircase which is symmetric about and includes the main diagonal.

A *staircase* (or *generalized block band*) matrix with s *steps* is a partial hermitian $n \times n$ matrix, R with precisely $s+1$ maximal specified hermitian submatrices, $\{R_1 \ \cdots \ R_{s+1}\}$, which are defined by

$$R_i = \begin{pmatrix} a_{j_i,j_i} & \cdots & a_{j_i,k_i} \\ \cdot & \cdots & \cdot \\ a_{k_i,j_i} & \cdots & a_{k_i,k_i} \end{pmatrix},$$

where $j_1 = 1$, $j_i < k_i$, $k_{s+1} = n$, and $j_i < j_{i+1} \le k_i + 1 \le k_{i+1} + 1$. The inertia of each of the R_i's is denoted by $\text{In}\, R_i = (\pi_i, \nu_i, \delta_i)$.

Staircase matrices allow the nondiagonal blocks to be nonsquare rectangles. Note that R_1 need not overlap R_3 as would be required in a block band matrix. In fact, R_1 need not even overlap R_2, but the main diagonal must be contained in the union of the R_i's. Also, the definition includes block *diagonal* matrices.

The next theorem shows that a staircase matrix, with all maximal submatrices being invertible, has hermitian completions with all the inertias consistent with Poincare's inequalities.

THEOREM 1.2 (an inertial triangle). *Given an s-step hermitian staircase $m \times m$ matrix R, suppose that each of the maximal submatrices $R_1, R_2, \ldots, R_s$ of R is invertible. Then the inertial polygon of R is the triangle*

$$\max\{\pi_i\} \le p, \qquad \max\{\nu_i\} \le n, \qquad p + n \le m.$$

The proof of Theorem 1.2 is presented in section 6, along with several theorems about the possible inertias of hermitian completions of staircase matrices. We will employ the method of Dym and Gohberg [DG], which decomposes the completion process into a succession of simple steps, each of which is a Theorem 1.1 step.

These results generalize the (scalar) band hermitian completion results of the second author in [D6]. A related result of Johnson and Rodman is restated as Lemma 5.4 herein.

We state our results for complex hermitian matrices, but they are all equally valid in the real symmetric case.

2. Preliminaries.

2.1. Notation. We shall denote by $\underline{p}, \underline{n}$ and $\overline{p}, \overline{n}$ the minimal and maximal, respectively, possible values of the positivity and the negativity of the completions of a given partially specified matrix.

Similarly, $\underline{r}$ and $\overline{r}$ will denote the minimal and maximal possible values for the rank of completion matrices of a given partially specified matrix. We have the obvious inequality $\overline{r} \le \overline{p} + \overline{n}$, which is generally strict. The determination of the maximal rank $\overline{r}$ for arbitrary (including nonband) hermitian completion problems is done in [CD]. In fact, it is shown there that the maximal completion rank does not increase if the assumption that the completion is hermitian is dropped; consequently, this rank can be computed explicitly using a result of [CJRW].

The inequality $\underline{r} \ge \underline{p} + \underline{n}$ is similarly obvious; however, it becomes an equality (i.e., $\underline{r} = \underline{p} + \underline{n}$) in many cases, including that of Theorem 1.1 (see Corollary 5.1) and certain block band matrices (see Theorem 6.2).

$J(p, n, d)$ will denote the square matrix $I_p \oplus -I_n \oplus 0_d$ of inertia (p, n, d). Sometimes we shall use the triple (π, ν, δ) to denote the inertia of a given (maximal) specified submatrix and (p, n, d) to denote the inertia of a hermitian completion of a given partial matrix. Congruence of matrices is denoted by $\cong$.

If a square matrix X is written in block form, say $X = (\, X_{ij},\ i, j = 1, \ldots, k\,)$ and X_{ij} is of size $a_i \times a_j$, we shall describe X as having block sizes $(a_1, \ldots, a_k)$.

2.2. Schur complements. Let $H = \left(\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right)$ If A is invertible then the *Schur complement* of A is

$$C^\times = C - B^* A^{-1} B. \tag{2.1}$$

Haynesworth [H] has shown that H is congruent to $A \oplus C^\times$. In particular,

$$\mathrm{In}\begin{pmatrix} A & B \\ B^* & C \end{pmatrix} = \mathrm{In}(A) + \mathrm{In}(C^\times). \tag{2.2}$$

More generally, if H is a $k \times k$ block matrix, and T is a subset of $\{1, \ldots, k\}$, let A be the principal submatrix whose block indices are in T. We can move A to the left upper corner by a permutation of the coordinates, and proceed as before, provided A is invertible. This procedure will be referred to as *complementation with respect to coordinates* T. The block division will always be clear from the context.

A similar procedure is available for non-hermitian matrices, yielding the weaker identity for $H = \left(\begin{smallmatrix} A & B \\ D & C \end{smallmatrix}\right)$ and $C^x = C - DA^{-1}B$:

$$\operatorname{rank}(H) = \operatorname{rank}(A) + \operatorname{rank}(C^{\times}). \tag{2.3}$$

We shall refer to this procedure as *non-hermitian complementation*.

2.3. Canonical forms. We shall repeatedly use the following terminology:

(i) *Equivalence canonical form:* It is well known that two matrices A and B are *equivalent* if there exist two invertible matrices S and T such that $A = SBT$. Every matrix A can be transformed by equivalence to the form $\left(\begin{smallmatrix} I_a & 0 \\ 0 & 0 \end{smallmatrix}\right)$, where $a = \operatorname{rank} A$.

We shall also need the following case of "block equivalence": For every matrix in block form $X = (A, B)$ of size $n \times (m_1 + m_2)$, there are invertible matrices S, T_1, and T_2 of size $n \times n$, $m_1 \times m_1$, and $m_2 \times m_2$, respectively, such that

$$S\begin{pmatrix} A & B \end{pmatrix}\begin{pmatrix} T_1 & 0 \\ 0 & T_2 \end{pmatrix} = \begin{pmatrix} I_a & 0 & 0 & 0 & 0 & 0 \\ 0 & I_b & 0 & 0 & I_b & 0 \\ 0 & 0 & 0 & 0 & 0 & I_c \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{2.4}$$

where

$$\begin{gathered} a = \operatorname{rank}(A, B) - \operatorname{rank}(B), \quad c = \operatorname{rank}(A, B) - \operatorname{rank}(A), \\ b = \operatorname{rank}(A) + \operatorname{rank}(B) - \operatorname{rank}(A, B). \end{gathered} \tag{2.5}$$

(ii) *Strong congruence canonical form:* Every hermitian matrix A of inertia (p, n, d) is congruent to a matrix of the form $J(p, n, d)$.

(iii) *Weak congruence canonical form:* Every hermitian matrix A of rank r is congruent to a matrix of the form $A' \oplus 0$, where A' is an invertible $r \times r$ matrix.

2.4. Restricted congruence. If P is a partial matrix, and S is invertible, the matrix $P' = S^*PS$ can be interpreted as a partial matrix in the following sense: an entry p'_{ij} of P' is determined if it is equal to $\{S^*HS\}_{ij}$ for every possible completion H of P. We call $P \to S^*PS$ *a restricted* congruence if p_{ij} being a specified entry of P implies that p'_{ij} is specified in P'.

There are some similarities between our concept of "restricted congruence" and Ball, Gohberg, Rodman, and Shalom's concept of "lower similarity" in [BGRS].

We will use restricted congruence in two ways:

1) block-diagonal congruence is used to put some (specified or unspecified) blocks of P in canonical form;

2) some row and column operations are used to annihilate blocks in P.

In some cases, unspecified blocks may become specified (in fact, annihilated) by congruence.

For example, the 1,1 block of $\left(\begin{smallmatrix} ? & I \\ I & 0 \end{smallmatrix}\right)$ can be annihilated by the process

$$\begin{pmatrix} I & -\frac{1}{2}Z \\ 0 & I \end{pmatrix}\begin{pmatrix} Z & I \\ I & 0 \end{pmatrix}\begin{pmatrix} I & 0 \\ -\frac{1}{2}Z & I \end{pmatrix} = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}. \tag{2.6}$$

Similarly,

$$\begin{pmatrix} Z & X & I \\ X^* & Y & 0 \\ I & 0 & 0 \end{pmatrix}$$

can be simplified to

$$\begin{pmatrix} 0 & 0 & I \\ 0 & Y & 0 \\ I & 0 & 0 \end{pmatrix}$$

as follows:

$$\begin{pmatrix} I & 0 & -\frac{1}{2}Z \\ 0 & I & -X^* \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} Z & X & I \\ X^* & Y & 0 \\ I & 0 & 0 \end{pmatrix} \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -\frac{1}{2}Z & -X & I \end{pmatrix} = \begin{pmatrix} 0 & 0 & I \\ 0 & Y & 0 \\ I & 0 & 0 \end{pmatrix}. \tag{2.7}$$

3. Inequalities: Necessary conditions. The shape of an inertial polygon for a hermitian completion problem is, to a large extent, determined by a few inequalities relating matrices and submatrices:

(i) *Rank interlacing.* If A is a $k \times l$ rectangular block of the $m \times m$ matrix H, we have

$$\operatorname{rank}(A) \le \operatorname{rank}(H) \le \operatorname{rank}(A) + (m-k) + (m-l). \tag{3.1}$$

(ii) *Poincaré inequalities.* Let A be a $k \times k$ principal submatrix of the $m \times m$ hermitian matrix H. Let λ_i and μ_i be the ordered eigenvalues of A and H, respectively. The Cauchy interlacing theorem states that $\lambda_i \le \mu_i \le \lambda_{i+m-k}$ (see, e.g., Theorem 4.3.15 in [HJ]). Equivalent statements are the Poincaré inequalities:

$$\pi(A) \le \pi(H) \le \pi(A) + m - k, \qquad \nu(A) \le \nu(H) \le \nu(A) + m - k. \tag{3.2}$$

The upper and lower bounds of (3.2) were strengthened in Theorem 1.2 of [D2]; the lower bound was strengthened as follows:

(iii) *Extended Poincaré's inequalities.*([D2]) Given a hermitian matrix in block form, $H = \left(\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right)$, set

$$\Delta = \operatorname{rank} \begin{pmatrix} A & B \end{pmatrix} - \operatorname{rank} A = \operatorname{Dim\ Ker} A - \operatorname{Dim\ Ker} \begin{pmatrix} A & B \end{pmatrix}^*.$$

Then

$$\operatorname{In} H \ge \operatorname{In} A + \Delta(1, 1, -1),$$

$$\pi(H) \ge \pi(A) + \Delta, \quad \text{and} \quad \nu(H) \ge \nu(A) + \Delta. \tag{3.3}$$

Inequalities (3.1), (3.2), (3.3) form a set of a priori bounds on completion inertias. In fact, in [CJRW] it is proved that the upper bound in (3.1) is sufficient for the determination of the maximal completion rank in the case of non-hermitian completions, and in [CD] the same is shown in the case of hermitian completions. It turns out that the necessary conditions of type (3.1), (3.2), (3.3) are also sufficient in determining the full inertial polygon in many cases, including the bordered matrix case (Theorem 1.1) and the block diagonal case [D1]. We shall emphasize cases of sufficiency of these conditions in the text.

4. Towards the 3×3 bordered case. This section paves the way for the analysis of the bordered case, which is carried out in section 5. The material in this section has independent value, and much of it is well known. We shall compute the inertial polygon for a 3×3 block pattern of the form

$$\begin{pmatrix} A & ? & ? \\ ? & B & ? \\ ? & ? & ? \end{pmatrix} \tag{4.1}$$

as Lemma 4.8. The special cases $\left(\begin{smallmatrix} A & ? \\ ? & ? \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} A & ? \\ ? & B \end{smallmatrix}\right)$ originally due to Cain and Sá, will also be reviewed as Lemmas 4.1 and 4.5. We shall also compute the possible inertias of a matrix of the form $A+X$ with inertia limitations on the unknown matrix X as Lemma 4.3. The results of Lemma 4.1 will be used to establish Lemmas 4.3 and 4.8 . The result of Lemma 4.3 will be used to establish Lemma 4.5 which in turn will be used in the proof of Lemma 4.8 which, in turn, will be used in the proof of Theorem 1.1. The proofs of Theorem 1.1 and Lemma 5.7 consist of reductions to the case of (4.1), which itself is of independent interest.

LEMMA 4.1. *Let* $H = \left(\begin{smallmatrix} H_1 & ? \\ ? & ? \end{smallmatrix}\right)$ *be a partially specified hermitian matrix of block sizes* (α, β). *Then the inertial polygon for* H *is the pentagon that contains all the lattice points* $(\pi(H), \nu(H))$ *which satisfy the inequalities*

$$\begin{array}{l} \pi(H_1) \le \pi(H) \le \pi(H_1) + \beta, \\ \nu(H_1) \le \nu(H) \le \nu(H_1) + \beta, \end{array} \qquad \pi(H) + \nu(H) \le \alpha + \beta. \tag{4.2}$$

Proof. The necessity of (4.2) follows from (3.1) and (3.2). For sufficiency, put H_1 in diagonal form, and complete H to a diagonal matrix. It is easy to show that every inertia in (4.2) is obtained. □

See also Theorem 1 in [S] and Theorem 1.2 in [D2].

COROLLARY 4.2. *In Lemma* 4.1 *the extremal values are*

$$\begin{array}{lll} \underline{p} = \pi(H_1), & \overline{p} = \pi(H_1) + \beta, & \underline{r} = \operatorname{rank} H, \\ \underline{n} = \nu(H_1), & \overline{n} = \nu(H_1) + \beta, & \overline{r} = \min\{\alpha + \beta, \overline{p} + \overline{n}\}. \end{array}$$

Moreover, H *in Lemma* 4.1 *admits positive definite, nonnegative definite, or invertible completions if and only if* H_1 *is positive definite, nonnegative definite, or* $\operatorname{rank}(H_1) \ge \alpha - \beta$, *respectively.*

The following result can be deduced with some effort from Theorem 2 in [S].

LEMMA 4.3. *Let* A *and* X *be* $m \times m$ *hermitian matrices. We consider* A *to be fixed and* X *to be a variable matrix with* $\pi(X) \le a$ *and* $\nu(X) \le b$. *Then the possible inertias of* $B = A + X$ *are the nonnegative lattice points satisfying the following inequalities:*

$$\begin{array}{rcccl} \pi(A) - b & \le & \pi(B) & \le & \pi(A) + a, \\ \nu(A) - a & \le & \nu(B) & \le & \nu(A) + b, \\ & & \operatorname{rank} B & \le & m. \end{array} \tag{4.3}$$

Proof. Necessity is obvious due to Sylvester's inertia principle. For sufficiency, take A to be diagonal, and restrict X to be diagonal as well. It is easy to show that every inertia in (4.3) is obtained. □

COROLLARY 4.4. *In Lemma* 4.3, $\underline{r} = \max\{\pi - b, 0\} + \max\{\nu - a, 0\}$. *Also, A admits positive definite completions if and only if* $\nu \le a$ *and* $\pi \ge m - a$ *, nonnegative definite completions if and only if* $\nu \le a$, *and invertible completions if and only if* $\delta \le a + b$.

For the values of $\underline{n}$, $\overline{n}$ in Lemma 4.3, see also [CG3, Lemma 2.2].

The following result is due to Cain and Sá.

LEMMA 4.5. ([CS]) *Let* $H = \left(\begin{smallmatrix} F & ? \\ ? & G \end{smallmatrix}\right)$ *be hermitian of block sizes* (α, γ). *Then the inertial polygon of* H *is determined by these inequalities:*

$$\begin{array}{ccccc} \max\{\pi(F), \pi(G)\} & \le & \pi(H) & \le & \min\{\pi(F) + \gamma, \pi(G) + \alpha\}, \\ \max\{\nu(F), \nu(G)\} & \le & \nu(H) & \le & \min\{\nu(F) + \gamma, \nu(G) + \alpha\}, \\ -\nu(F) - \nu(G) & \le & \pi(H) - \nu(H) & \le & -\pi(F) - \pi(G), \\ & & \mathrm{rank} H & \le & \alpha + \gamma \ . \end{array} \tag{4.4}$$

A generalization of Lemma 4.5 and (4.4) to more than two diagonal blocks can be found in Brian Cain's paper [C]. A short proof of the necessity part of Cain's result was obtained by J. Dancis ([D1, Corollary 11.1 and Lemma 11.2]). In the 2×2 block case, this proof is given below.

Proof of necessity of inequalities (4.4). Let$H(X) = \left(\begin{smallmatrix} F & X \\ X^* & G \end{smallmatrix}\right)$ be a completion of H. Set

$$t = \dim\left(\ker\begin{pmatrix} F \\ X^* \end{pmatrix} \cap \ker F\right), \qquad t' = \dim\left(\ker\begin{pmatrix} X \\ G \end{pmatrix} \cap \ker\ G\right).$$

Noting that

$$\mathrm{rank}\begin{pmatrix} F \\ X^* \end{pmatrix} - \mathrm{rank} F = \delta(F) - t, \qquad \delta(H(X)) \ge t + t',$$

we apply the extended Poincare inequalities (3.3) to both F and G and we obtain

$$2\nu(H(X)) \ge \nu(F) + \nu(G) + \delta(F) - t + \delta(G) - t'$$

$$\ge \nu(F) + \nu(G) + \delta(F) + \delta(G) - \delta(H(X)) \ .$$

Subtracting this inequality from $\nu(H) + \pi(H) = \alpha + \gamma - \delta(H)$ will yield the right side of (4.4). A symmetric argument produces the left inequality. □

Proof of Lemma 4.5. We show by reduction to Lemma 4.3 that inequalities (4.4) describe the inertial polygon. Let $H(X) = \left(\begin{smallmatrix} F & X \\ X^* & G \end{smallmatrix}\right)$ be a completion of H. Putting F in weak canonical form, and taking the Schur complement of the new first coordinate F', we calculate, using (2.2),

$$H(X) \cong \begin{pmatrix} F' & 0 & X_1 \\ 0 & 0 & X_2 \\ X_1^* & X_2^* & G \end{pmatrix} \cong F' \oplus H', \qquad H' = \begin{pmatrix} 0 & X_2 \\ X_2^* & G^\times \end{pmatrix},$$

where $G^\times = G - X_1^* F'^{-1} X_1$. Hence

$$\mathrm{In}(H(X)) = \mathrm{In}(H') + (\pi(F), \nu(F), 0). \tag{4.5}$$

Putting $G^{\times}$ in weak canonical form, we get

$$H' \cong \begin{pmatrix} 0 & X_3 & X_4 \\ X_3^* & G^{\times\prime} & 0 \\ X_4^* & 0 & 0 \end{pmatrix}.$$

Putting X_4 in equivalence canonical form, we get

$$H' \cong \begin{pmatrix} 0 & 0 & X_5 & I_r & 0 \\ 0 & 0 & X_6 & 0 & 0 \\ X_5^* & X_6^* & G^{\times\prime} & 0 & 0 \\ I_r & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

We set $V'' = -X_6 G^{\times\prime -1} X_6^*$, and denote

$$\text{In}(G^{\times}) = (\pi'', \nu'', \delta''); \qquad \text{In}(V'') = (\hat{\pi}, \hat{\nu}, \hat{\delta}); \qquad \text{rank}(X_4) = r. \tag{4.6}$$

Removing coordinate 5 and taking the Schur complement of coordinates 1, 3, and 4, in H', we calculate using (2.2) and (2.7)

$$H' \cong J(r + \pi'', r + \nu'', *) \oplus V''. \tag{4.7}$$

We now develop the relevant inequalities involving the dummy variables in (4.6). By Lemma 4.3 the inertial polygon of $G^{\times}$ is determined by the inequalities on (π'', ν'') :

$$\begin{array}{ll} -\nu(F) \le \pi'' - \pi(G) \le \pi(F), & 0 \le \pi'', \quad 0 \le \nu'', \\ -\pi(F) \le \nu'' - \nu(G) \le \nu(F), & \pi'' + \nu'' \le m_2. \end{array} \tag{4.8}$$

By Lemma 4.3 again, we compute the inertial polygon for V'':

$$0 \le \hat{\pi} \le \nu'', \quad 0 \le \hat{\nu} \le \pi'', \quad \hat{\pi} + \hat{\nu} \le m_1 - \text{rank } F - r. \tag{4.9}$$

The only restriction on r is the size of X_4 :

$$0 \le r \le m_2 - \pi'' - \nu''. \tag{4.10}$$

Now (4.4) is obtained from (4.5)–(4.9) by eliminating π'', ν'', $\hat{\pi}$, $\hat{\nu}$, and r. □

The values of $\underline{p}$, $\overline{p}$, $\underline{n}$, $\overline{n}$ can easily be computed from Lemma 4.5. The values $\underline{n}$ and $\overline{n}$ were computed in [CS].

LEMMA 4.6. (See [D1].) *In Lemma* 4.5 *we have*

$$\underline{p} = \max\{\pi(F), \pi(G)\}, \qquad \underline{n} = \max\{\nu(F), \nu(G)\}, \qquad \underline{r} = \underline{p} + \underline{n}. \tag{4.11}$$

Moreover, there exists a matrix X *which simultaneously achieves the minimal possible* ranks *for* $\left(\begin{smallmatrix} F & X \\ X^* & G \end{smallmatrix}\right)$ $\left(\begin{smallmatrix} F \\ X^* \end{smallmatrix}\right)$, *and* $\left(\begin{smallmatrix} X \\ G \end{smallmatrix}\right)$, *namely,*

$$\text{rank}\begin{pmatrix} F & X \\ X^* & G \end{pmatrix} = \max\{\pi(F), \pi(G)\} + max\{\nu(F), \nu(G)\},$$

$$\text{rank}\begin{pmatrix} F & X \end{pmatrix} = \text{rank}(F), \quad \textit{and} \quad \text{rank}\begin{pmatrix} X \\ G \end{pmatrix} = \text{rank}(G).$$

Proof. The values of $\underline{p}$ and $\underline{n}$ follow directly from Lemma 4.5. To show the rest, we may put F and G in strong canonical form:

$$F = \begin{pmatrix} I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad X = \begin{pmatrix} X_{1,1} & X_{1,2} & X_{1,3} \\ X_{2,1} & X_{2,2} & X_{2,3} \\ X_{3,1} & X_{3,2} & X_{3,3} \end{pmatrix}.$$

Choose the diagonal entries of $X_{1,1}$ and $X_{2,2}$ by the rule $(X_{1,1})_{ii} = 1$ and $(X_{2,2})_{ii} = 1$. Choose all other entries of X to be zero. We get a completion with the desired minimal ranks. □

The original result of J. Dancis ([D1, Theorem 1.3]) is in fact more general in two respects: first, it extends to more than two block diagonals. Moreover, it is not restricted to *minimal* ranks: it shows, more generally, that any choice of kernels of a column decomposition as well as any choice of inertia which is consistent with the extended Poincaré inequalities can be obtained.

THEOREM 4.7 (a constrained hermitian completion (see [D1])). *Given hermitian matrices H_{ii}, $i = 1, \dots, s$, with inertias* $\operatorname{In} H_{ii} = (\pi_i, \nu_i, \delta_i)$ *and sizes $n_i = \pi_i + \nu_i + \delta_i$, let $S = \oplus H_{ii}$ be the block diagonal matrix of size $n = \Sigma n_i$. Choose a subspace $K_i \subset \ker H_{ii}$ such that* $\dim \ker H_{ii} - \dim K_i \le n - n_i$. *Let $r_i = n_i - \dim \ker K_i$. Set*

$$\Delta_i = r_i - \operatorname{Rank} H_{ii} \quad for \ \ each \ \ i = 1, 2, \dots, s.$$

Then an integer triple (π, ν, δ) satisfying the equality $\pi + \nu + \delta = n$ is the inertia of a hermitian completion H of S with column block structure:

$$H = (M_1, M_2, \dots, M_s), \ \ where \ \ each \ \ M_i \ \ is \ \ an \ \ n \times n_i \ matrix, \ \ and \ \ \ker M_i = K_i$$

if and only if (π, ν, δ) satisfies the inequalities

$$\pi \ge \operatorname{Max}\{\pi_i + \Delta_i\}, \quad \nu \ge \operatorname{Max}\{\nu_i + \Delta_i\}, \quad and \quad \delta \ge \Sigma(\delta_i - \Delta_i).$$

(The notation here is different than the one used in [D1]: the r_i and Δ_i here correspond to $\pi_i + \nu_i + \delta_i - d_i$ and $\delta_i - d_i$ of [D1], respectively.)

The next lemma is the main result of this section; it combines Lemmas 4.1 and 4.5.

LEMMA 4.8. *Let P be a partial matrix of the form*

$$\begin{pmatrix} F & ? & ? \\ ? & G & ? \\ ? & ? & ? \end{pmatrix}$$

and of block sizes $(\alpha, \gamma, \epsilon)$. Then the inertial polygon of P consists of the lattice points determined by these inequalities:

$$\begin{array}{rcccl} \max\{\pi(F), \pi(G)\} & \le & \pi(P) & \le & \epsilon + \min\{\pi(F) + \gamma, \pi(G) + \alpha\}, \\ \max\{\nu(F), \nu(G)\} & \le & \nu(P) & \le & \epsilon + \min\{\nu(F) + \gamma, \nu(G) + \alpha\}, \\ -\epsilon - \nu(F) - \nu(G) & \le & \pi(P) - \nu(P) & \le & \epsilon + \pi(F) + \pi(G), \\ \pi(P) + \nu(P) & = & \operatorname{rank} P & \le & \alpha + \gamma + \epsilon \ . \end{array} \tag{4.12}$$

The sufficiency proof of inequalities (4.12) is the same as for inequalities (4.4).

Proof. Every completion H of P has the form $\left(\begin{smallmatrix} H_1 & * \\ * & * \end{smallmatrix}\right)$, where $\left(\begin{smallmatrix} F & * \\ * & G \end{smallmatrix}\right)$. $\operatorname{In}(H_1)$ was computed in Lemma 4.5. By Lemma 4.1, $\operatorname{In}(H_1)$ and $\operatorname{In}(H)$ are connected by

$$0 \le \pi(H) - \pi(H_1) \le \epsilon, \qquad 0 \le \nu(H) - \nu(H_1) \le \epsilon, \tag{4.13}$$

and eliminating $\mathrm{In}(H_1)$ from inequalities (4.4) and (4.13), and using the identities

$$\alpha = \pi(F) + \nu(F) + \delta(F), \qquad \gamma = \pi(G) + \nu(G) + \delta(G),$$

we get inequalities (4.12). □

5. Inertias of block bordered matrices. In this section we establish Theorem 1.1 using the results stated in sections 3 and 4. The material in this section is new. The scalar case was classified in [D3]. Other special cases of Theorem 1.1 occur in [D1], [L], [G], and [CG3].

Sections 5.3–5.5 contain additional results and corollaries of Theorem 1.1 concerning minimal rank completions of various types, and the case where the two maximal specified hermitian submatrices R_1 and R_2 of $P(Z)$ of Theorem 1.1 are invertible. In section 5.6 we present a small application to the algebraic matrix Riccati equation $A + AZ^* + ZB^* + ZCZ^* = 0$: a criterion for solvability and a characterization of the possible inertias of the solution matrix Z (which need not be hermitian).

5.1. Internal relations for bordered matrices. For the bordered matrix $P(Z)$ of Theorem 1.1, we note that $R_1 = \left(\begin{smallmatrix} A & B \\ B^* & C \end{smallmatrix}\right)$ and $R_2 = \left(\begin{smallmatrix} C & D \\ D^* & E \end{smallmatrix}\right)$ are the maximal specified hermitian submatrices of $P(Z)$ and $Q = [B^*,\ C,\ D]$ is the maximal specified non-hermitian submatrix of $P(Z)$.

Observation 5.1 (internal relations for a bordered matrix). With the notation of Theorem 1.1, we define

$$\hat{\Delta} = \max\{\Delta', \Delta''\}$$

and

$$d' = \text{rank } [B^*, C] - \text{rank } C, \quad d'' = \text{rank } [C, D] - \text{rank } C, \quad \hat{d} = \max\{d', d''\}.$$

Then

$$d' - \Delta'' = d'' - \Delta' \ge 0, \tag{5.1}$$

$$\nu \ge \nu(C) + d', \qquad \pi' \ge \pi(C) + d'', \tag{5.2}$$

$$\text{rank } R_1 \ge \text{rank } C + 2\hat{d}, \tag{5.3}$$

$$r = \text{rank } C + d' + \Delta' = \text{rank } C + d'' + \Delta''. \tag{5.4}$$

Proof. The inequality of (5.1) follows from rank considerations. The equality, as well as (5.4), follows from the definitions. Applying the extended Poincaré's inequalities to C as a submatrix of R_1 *or* R_2 provides (5.2). Equation (5.3) also follows from the extended Poincaré's inequalities. □

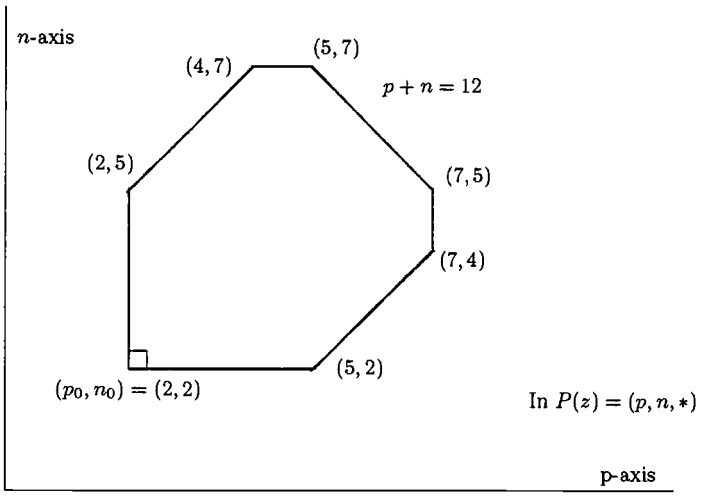


FIG. 5.1. *A seven-sided inertial polygon*

5.2. Proof of Theorem 1.1. Before proving the theorem, let us comment on the necessity and minimality of its conditions:

Minimality. The inertial polygon for a simple matrix, with all seven edges present, is illustrated in Fig. 5.1. The matrix chosen was of block sizes $(6, 1, 6)$ with $A = I_2 \oplus -I_2 \oplus 0_2$, $E = 1 \oplus -1 \oplus 0_4$, and B, C, and D are zero matrices of appropriate order. This results in r and the three Δ's being zero. This example shows that the set of seven inequalities defining the inertial diamond is not redundant.

Necessity. Necessity of each one of the seven inequalities can be easily demonstrated: inequality (1.4) follows from (3.1). The upper bounds in inequalities (1.1)–(1.2) are a consequence of (3.2). The lower bounds in (1.1) and (1.2) are just the extended Poincaré's inequalities (3.3). It remains to derive (1.3).

CLAIM 5.2. *The extended Poincaré's inequalities* (3.3) *imply* (1.3).

Proof. We define the partial matrices

$$N' = \begin{pmatrix} A & B & ? \\ B^* & C & D \end{pmatrix}, \qquad N'' = \begin{pmatrix} B^* & C & D \\ ? & D^* & E \end{pmatrix}, \tag{5.5}$$

and their completions

$$N'(Z) = \begin{pmatrix} A & B & Z \\ B^* & C & D \end{pmatrix}, \qquad N''(Z) = \begin{pmatrix} B^* & C & D \\ Z & D^* & E \end{pmatrix}. \tag{5.6}$$

Let

$$\Gamma = \text{rank } N' - \text{rank } R_1 \quad \text{and} \quad \Gamma' = \text{rank } N'' - \text{rank } R_2.$$

Using the extended Poincaré's inequalities(3.3) twice, we have

$$2p \geq \pi + \pi' + \Gamma + \Gamma'.$$

Substituting $p = \alpha + \beta + \gamma - n - d$ for one of the p's on the left-hand side, we obtain

$$p - n \geq \pi + \pi' + \Gamma + \Gamma' + d - \alpha - \beta - \gamma. \tag{5.7}$$

Using the identities $\pi + \nu + \delta = \alpha + \beta$ and $\pi' - \beta - \gamma = -\nu' - \delta'$ and (5.7) we obtain

$$p - n \geq -\nu - \nu' - \delta - \delta' + \Gamma + \Gamma' + d + \beta. \tag{5.8}$$

We note that

$$\dim \ker P \geq \dim \ker N' + \dim \ker N'' - \dim \ker N' \cap N''.$$

But this translates into

$$d \geq (\delta - \Gamma) + (\delta' - \Gamma') - (\beta - r),$$

or equivalently

$$-r - \delta - \delta' + \Gamma + \Gamma' + d + \beta \geq 0. \tag{5.9}$$

Finally, (5.8) and (5.9) establish (1.3). □

We will establish Theorem 1.1 by using Schur complements (equation (2.2)) and row and column operations (equation(2.7)) and the other forms presented in section 2, repeatedly, in order to reduce Theorem 1.1 to Lemma 4.8.

Proof of Theorem 1.1. We begin by putting C in weak canonical form:

$$P(Z) = \begin{pmatrix} A & B' & B" & Z \\ B'^* & C' & 0 & D' \\ B"^* & 0 & 0 & D" \\ Z^* & D'^* & D"^* & E \end{pmatrix}, \tag{5.10}$$

with block sizes $(\alpha, \operatorname{rank} C, \beta - \operatorname{rank} C, \gamma)$. Taking the Schur complement of C' in $P(Z)$ as in (2.2) yields

$$\operatorname{In} P(Z) = \operatorname{In} \hat{H} + \operatorname{In} C', \tag{5.11}$$

where

$$\hat{H} = \begin{pmatrix} F & B" & Y \\ B"^* & 0 & D" \\ Y^* & D"^* & G \end{pmatrix}, \qquad \begin{aligned} F &= A - B'C'^{-1}B'^*, \\ G &= E - D'^*C'^{-1}D', \\ Y &= Z - B'C'^{-1}D'. \end{aligned}$$

Next we put $[B"^*, D"]$ in the canonical form (2.4). Using (5.10), we obtain the matrix

$$\hat{H} = \begin{pmatrix} F_{11} & F_{12} & F_{13} & I_{\Delta''} & 0 & 0 & 0 & X_{11} & X_{12} & X_{13} \\ F_{12}^* & F_{22} & F_{23} & 0 & I_b & 0 & 0 & X_{21} & X_{22} & X_{23} \\ F_{13}^* & F_{23}^* & F_{33} & 0 & 0 & 0 & 0 & X_{31} & X_{32} & X_{33} \\ I_{\Delta''} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_b & 0 & 0 & 0 & 0 & 0 & 0 & I_b & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{\Delta'} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ X_{11}^* & X_{21}^* & X_{31}^* & 0 & 0 & 0 & 0 & G_{11} & G_{12} & G_{13} \\ X_{12}^* & X_{22}^* & X_{32}^* & 0 & I_b & 0 & 0 & G_{12}^* & G_{22} & G_{23} \\ X_{13}^* & X_{23}^* & X_{33}^* & 0 & 0 & I_{\Delta'} & 0 & G_{13}^* & G_{23}^* & G_{33} \end{pmatrix}. \tag{5.12}$$

The block sizes here are $(\Delta'', b, \alpha', \Delta'', b, \Delta', \beta', \gamma', b, \Delta')$, where $b = d' - \Delta''$ and the X_{ij}'s are blocks of Y, the F_{ij}'s and G_{ij}'s are conforming blocks of F and G. The new block sizes are related to α, β, γ via

$$\alpha = \alpha' + d', \qquad \beta = \operatorname{rank} C + \beta' + d' + \Delta', \qquad \gamma = \gamma' + d''. \tag{5.13}$$

Next we use restricted congruence, see section 2. By row and column operations based on $H_{41} = H_{14} = I_{\Delta''}$ and $H_{10,6} = H_{6,10} = I_{\Delta'}$, we may assume without loss of generality that $F_{11}, F_{12}, F_{13}, X_{11}, X_{12}, X_{13}, X_{23}$, X_{33}, G_{13}, G_{23}, and G_{33} are all zero. This modifies the matrix $\hat{H}$, without changing its inertia, to

$$H' = \begin{pmatrix} 0 & 0 & 0 & I_{\Delta''} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & F_{22} & F_{23} & 0 & I_b & 0 & 0 & X_{21} & X_{22} & 0 \\ 0 & F_{23}^* & F_{33} & 0 & 0 & 0 & 0 & X_{31} & X_{32} & 0 \\ I_{\Delta''} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_b & 0 & 0 & 0 & 0 & 0 & 0 & I_b & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{\Delta'} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & X_{21}^* & X_{31}^* & 0 & 0 & 0 & 0 & G_{11} & G_{12} & 0 \\ 0 & X_{22}^* & X_{32}^* & 0 & I_b & 0 & 0 & G_{12}^* & G_{22} & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{\Delta'} & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.14}$$

We may discard row 7 and column 7, which are all zero. Next we complement H' with respect to the block

$$H'_{[1,2,4,5,6,10]} = \begin{pmatrix} 0 & 0 & I_{\Delta''} & 0 & 0 & 0 \\ 0 & F_{22} & 0 & I_b & 0 & 0 \\ I_{\Delta''} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_b & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{\Delta'} \\ 0 & 0 & 0 & 0 & I_{\Delta'} & 0 \end{pmatrix}.$$

The Schur complement turns out to be

$$H'' = \begin{pmatrix} F_{33} & X_{31} & X_{32} \\ X_{31}^* & G_{11} & G_{12} \\ X_{32}^* & G_{12}^* & G_{22} \end{pmatrix}$$

$$- \begin{pmatrix} 0 & F_{23}^* & 0 & 0 & 0 & 0 \\ 0 & X_{21}^* & 0 & 0 & 0 & 0 \\ 0 & X_{22}^* & 0 & I_b & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & I_{\Delta''} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_b & 0 & 0 \\ I_{\Delta''} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_b & 0 & -F_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{\Delta'} \\ 0 & 0 & 0 & 0 & I_{\Delta'} & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ F_{23} & X_{21} & X_{22} \\ 0 & 0 & 0 \\ 0 & 0 & I_b \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} F_{33} & X_{31} & X_{32} - F_{23}^* \\ X_{31}^* & G_{11} & G_{12} - X_{21}^* \\ X_{32}^* - F_{23} & G_{12}^* - X_{21} & G_{22} - X_{22} - X_{22}^* + F_{22} \end{pmatrix} = \begin{pmatrix} F_{33} & ? & ? \\ ? & G_{11} & ? \\ ? & ? & ? \end{pmatrix}, \tag{5.15}$$

with block sizes (α', γ', b) . By the Schur complement inertia formula (2.2), we get

$$\begin{aligned} \operatorname{In} H' &= \operatorname{In} H'_{[1,2,4,5,6,10]} + \operatorname{In} H'' + (0, 0, \beta) \\ &= \operatorname{In} H'' + (d' + \Delta', d' + \Delta', \beta'). \end{aligned} \tag{5.16}$$

(The β' accounts for removing coordinate 7 from H'.) Thus we have

$$\text{In } P(Z) = \text{In } C' + \text{In } H'' + (d' + \Delta', d' + \Delta', \beta'). \tag{5.17}$$

We will use (5.17) and Lemma 4.8 to calculate In $P(Z)$; to find In H'', we must first calculate In F_{33} and In G_{11} :

Claim 5.3.

$$\begin{aligned} \text{In } F_{33} &= \text{In } R_1 - (d', d', *) - \text{In } C', \\ \text{In } G_{11} &= \text{In } R_2 - (d'', d'', *) - \text{In } C'. \end{aligned} \tag{5.18}$$

Proof. By restricting (5.10)–(5.14) to the upper left corner R_1, we may take the Schur complement of C' as a submatrix of R_1; this yields

$$\text{In } R_1 = \text{In}\begin{pmatrix} A & B' & B" \\ B'^* & C' & 0 \\ B"^* & 0 & 0 \end{pmatrix} = \text{In } C' + \text{In}\begin{pmatrix} F & B" \\ B"^* & 0 \end{pmatrix}, \tag{5.19}$$

where F is as in (5.11). Using elimination, (2.7), we note that

$$\text{In}\begin{pmatrix} F & B" \\ B"^* & 0 \end{pmatrix} = \text{In}\begin{pmatrix} 0 & 0 & 0 & I_{\Delta''} & 0 & 0 \\ 0 & 0 & 0 & 0 & I_b & 0 \\ 0 & 0 & F_{33} & 0 & 0 & 0 \\ I_{\Delta''} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_b & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \text{In } F_{33} + (d', d', *), \tag{5.20}$$

where F_{33} is as in (5.12). Solving for In F_{33} in (5.19) and (5.20) yields the first part of (5.18). A similar argument holds for the second part. ∎

We proceed with the proof of Theorem 1.1. Applying the size identities

$$\alpha' = \pi(F_{33}) + \nu(F_{33}) + \delta(F_{33}), \qquad \gamma' = \pi(G_{11}) + \nu(G_{11}) + \delta(G_{11}),$$

and Lemma 4.8 (with its submatrices F and G corresponding to F_{33} and G_{11} here), we obtain these inequalities for In H'':

$$\begin{aligned} \max\{\pi(F_{33}), \pi(G_{11})\} \le \pi(H'') &\le d' - \Delta'' + \min\{\pi(F_{33}) + \gamma', \pi(G_{11}) + \alpha'\}, \\ \max\{\nu(F_{33}), \nu(G_{11})\} \le \nu(H'') &\le d' - \Delta'' + \min\{\nu(F_{33}) + \gamma', \nu(G_{11}) + \alpha'\}, \\ -(d' - \Delta'') - \nu(F_{33}) - \nu(G_{11}) \le \pi(H'') - \nu(H'') &\le d' - \Delta'' + \pi(F_{33}) + \pi(G_{11}), \\ \pi(H'') + \nu(H'') &\le \alpha' + \gamma' + d' - \Delta''. \end{aligned} \tag{5.21}$$

Inequalities (1.1)–(1.4) are obtained by plugging inequalities (5.21) and equation (5.18) into equation (5.17) and then using equations (5.1), (5.4), and (5.13) to eliminate all the intermediary inertias. ∎

5.3. Extremal inertia values and inertia-preserving completions. In this section and the next, we use the geometry of the inertial polygon as the basis (i) for establishing a minimum rank completion for the bordered matrix P of Theorem 1.1 (Corollary 5.2); and (ii) for showing that, assuming invertibility of R_1 and R_2, all the inertias which are consistent with Poincare's inequalities can be obtained by completion (Corollary 5.5). In section 6 we will use these results as building blocks for our proofs of completion theorems for "staircase" matrices.

First we show, under the notation of Theorem 1.1, that there is always a completion whose positivity and negativity are the minimal ones allowed by the extended Poincaré's inequalities. This implies that the minimal rank is $\underline{r} = \underline{p} + \underline{n}$.

COROLLARY 5.1 (minimal rank completions). *With the notation of Theorem* 1.1, *there exists a hermitian completion* $P = P(Z)$ *such that*

$$\pi(P(Z)) = p_0, \qquad \nu(P(Z)) = n_0.$$

Proof. We argue by inspection on Figure 5.1. If the vertex $v_0 = (p_0, n_0)$ is not in the inertial polygon for $P(z)$, this vertex must be cut off by one of the extremal lines defining inequalities (1.2)–(1.4). Since these inequalities are consistent, it is clear that (p_0, n_0) must satisfy (1.2) and (1.4). It remains to check the two inequalities (1.3). We have to consider four possible choices for p_0 and n_0 in the notation of Theorem 1.1:

(I) Suppose that

$$(p_0, n_0) = (\pi, \nu) + \Delta'(1, 1). \tag{5.22}$$

Rank considerations imply that Rank $C + d' + d'' \geq r$. Equation (5.1) implies that $\pi' + \nu \geq$ Rank $C + d' + d''$. Hence $-\nu \leq +\pi' - r$. This and (5.22) imply that $p_0 - n_0 = \pi - \nu \leq \pi + \pi' - r$, proving the right inequality in (1.3). Interchanging the roles of the $\pi' s$ and the $\nu' s$ establishes the left inequality.

(II) A similar proof applies to the case $(p_0, n_0) = (\pi, \nu) + \Delta''(1, 1)$.

(III) If

$$p_0 = \pi' + \Delta'' \quad \textit{and} \quad n_0 = \nu + \Delta', \tag{5.23}$$

then a combination of (5.1), (5.3), and (5.4) yields rank $R_1 \geq r + \Delta'' - \Delta'$. By simple algebra we get $\pi' - \nu + \Delta'' - \Delta' \leq \pi + \pi' - r$. This and (5.23) imply that $p_0 - n_0 = \pi' - \nu + \Delta'' - \Delta' \leq \pi + \pi' - r$, proving the right inequality in (1.3). Interchanging the roles of the $\pi' s$ and the $\nu' s$ establishes the left inequality.

(IV) A similar proof applies to the case $p_0 = \pi + \Delta'$ *and* $n_0 = \nu' + \Delta''$. □

Remark. Corollary 5.1 implies that the minimal rank of the set of hermitian completions of P is $\underline{r} = p_0 + n_0$.

Constantinescu and Gheondea presented in [CG3] another formula for $\underline{n}$.

Of particular interest is the case where the minimal rank solutions inherit their inertia values p and n from the specified blocks R_1 and R_2, i.e.,

$$p_0 = \max\{\pi_i\}, \qquad n_0 = \max\{\nu_i\}. \tag{5.24}$$

We call such completions *inertia preserving*. Note that (5.24) does not guarantee that the minimal completion rank is $\max\{\text{rank}R_1, \text{rank}R_2\}$. The following simple result will play a major role in finding inertia-preserving completions for block band matrices in section 6.

COROLLARY 5.2 (inertia preserving completions). *Assume the notation of Theorem* 1.1. *Suppose that* P *satisfies an "equality of* ranks*" condition*

$$r = \text{rank } [B^*, C] = \text{rank } [C, D]. \tag{5.25}$$

Then P *admits inertia-preserving completions.*

Indeed, under condition (5.25), the formulas of the inertial polygon simplify, since we have $\Delta' = \Delta'' = \hat{\Delta} = 0$. In particular, (5.24) holds. □

Condition (5.24) and Corollary 5.2 are implied by the stronger condition

$$r = \text{rank } C. \tag{5.26}$$

In the notation of Observation 5.1, (5.26) is equivalent to any of the following:

$$\text{rank } C = \text{rank}[B^*, C] = \text{rank}[C, D], \qquad d' = d'' = 0, \qquad \hat{d} = 0.$$

This condition is satisfied if, e.g., C is invertible.

Condition (5.25) is not necessary for the existence of inertia-preserving completions. As an example, consider the partial matrix

$$P = \begin{pmatrix} 1 & 1 & ? \\ 1 & 0 & 0 \\ ? & 0 & 0 \end{pmatrix}.$$

A simple argument, using the extended Poincaré's inequalities (3.3), shows that if (5.25) is not satisfied, then the inertia-preserving completion must inherit both p and n from the same block. Hence $\underline{r} = \max\{\text{rank } R_1, \text{rank } R_2\}$. In the above example, $\underline{r} = \text{rank } R_1$.

5.4. The width of the inertial set. We define the *width* w of the inertial polygon to be the maximal value of $|(p-n) - (p'-n')|$ over all pairs of points (p,n) and (p',n') belonging to the inertial polygon. See Fig. 1.1. This width equals the sum of the lengths of the two perpendicular sides of the inertial polygon. It is clear that inequality (1.3) puts a limitation on the width, namely, w cannot exceed the modulus of the difference of the right-and left-hand sides in this inequality:

$$w \le \text{rank } R_1 + \text{rank } R_2 - 2r. \tag{5.27}$$

(It can be shown directly that this value is always nonnegative.)

The sides, with slope of minus 1, come from inequality (1.3), which may be rewritten as

$$\pi - \nu' - (\text{rank } R_1 - r) \le p - n \le \pi - \nu' + (\text{rank } R_2 - r). \tag{5.28}$$

In this way

$$w \le (\text{rank } R_1 - r) + (\text{rank } R_2 - r) \tag{5.29}$$

is related to the width of the inertial polygon. The width of the inertial polygon tends to increase as we increase the ranks of R_1 and R_2. In this section we study the two extreme cases. The "slim" case is when Rank R_1 = Rank R_2 = Rank C. Here the polygon degenerates to a segment with a 45 degree inclination. The "fat" case occurs under the maximal rank condition det R_1 det $R_2 \neq 0$; here the polygon extends to maximum capacity, and fills a triangle (Corollary 5.5). We start with the "slim" case.

COROLLARY 5.3. *Given the notation of Theorem* 1.1*, suppose that*

$$\text{Rank } R_1 = \text{Rank } R_2 = \text{Rank } C. \tag{5.30}$$

Then the inertial polygon coincides with the segment

$$(p,n) = (\pi + k, \nu + k), \qquad k = 0, ..., \min\{\alpha, \gamma\}.$$

Moreover, the minimal rank *completion is unique.*

Proof. Condition (5.30) together with the Poincaré inequalities imply that C, R_1, and R_2 all have the same inertia $(\pi, \nu, *)$ (with possibly different nullities). The same condition also implies (5.26), hence (5.25), and Corollary 5.2 can be used. We conclude that

$$r = \text{rank } R_1 = \text{rank } R_2, \qquad p_0 = \pi, \quad n_0 = \nu. \tag{5.31}$$

That condition (5.31) implies zero width is clear from (5.27). The rest restricts the polygon to a line segment of the form $(\pi + k, \nu + k), \quad k = 0, \ldots, K$. The value $K = \min\{\alpha, \gamma\}$ follows from Theorem 1.1 with some algebra. It can also be deduced from the maximal rank considerations in [CD].

To prove uniqueness of the minimal rank completion, we note that (5.30) forces the factorizations $R_1 = \begin{pmatrix} S \\ I \end{pmatrix} C(S^* \;\; I), \quad R_2 = \begin{pmatrix} I \\ T \end{pmatrix} C(I \;\; T^*)$. Setting $Z = SCT^* + Z'$, one can check that the completion rank is rank $C + 2$ rank Z'. So the unique solution requires that $Z' = 0$. ∎

Observation 5.3 represents the extreme case of a "slim" inertial set. We now turn to examine the other extreme case of a "fat" inertial set. Under the assumption that R_1 and R_2 are invertible matrices, four inequalities among (1.1)–(1.4) are redundant, and the inertial polygon becomes a triangle, admitting any inertia compatible with the Poincaré inequalities and the size limitation. First we quote the following known result about matrices with chordal graphs. Chordality is discussed in [GJSW], [JR1], and [JR2], and it suffices to say that block bordered 3×3 patterns (and in fact the general staircase patterns of section 6) have chordal graphs.

LEMMA 5.4 (Corollary 6 of [JR1]). *In any hermitian partial matrix P of size $m \times m$ whose pattern has a chordal graph and all its maximal hermitian specified submatrices are invertible, the points $v_1 = (\underline{p}, m - \underline{p}, 0)$ and $v_2 = (m - \underline{n}, \underline{n}, 0)$, together with all the lattice points on the straight line segment connecting them, belong to the inertial set of P.*

In the bordered case we can say more.

COROLLARY 5.5. *Assume the notation of Theorem* 1.1. *Suppose that R_1 and R_2 are invertible. Then the inertial polygon is the triangle whose vertices are*

$$v_0 = (p_0, n_0), \qquad v_1 = (p_0, \alpha + \beta + \gamma - p_0), \qquad v_2 = (\alpha + \beta + \gamma - n_0, n_0).$$

In other words, every inertia consistent with the Poincaré inequalities $p_0 \le p$, $n_0 \le n$, and $p + n \le \alpha + \beta + \gamma$ is allowed.

Proof. Let T be the triangle defined by the above three inequalities. Let D be the inertial polygon. It is easy to check that v_0, v_1, v_2 are the three vertices of T. Since the Poincaré inequalities are a subset of (1.1)–(1.4), we get the inclusion $D \subset T$. Note that in (1.4) our hypothesis implies that $r = \beta$.

On the other hand, $v_0 \in D$ by Corollary 5.1, and $v_1, v_2 \in D$ by Lemma 5.4. By convexity, we conclude that $T \subset D$. ∎

5.5. Simultaneous rank minimization. We now strengthen the minimal rank result obtained in the last section (Corollary 5.1). Consider the partial matrices N' and N'' of (5.5). We wish to find a matrix Z which will simultaneously induce minimal rank completions in N' and N'' as well as in the full bordered matrix P.

Before we tackle the general case, let us make the simplifying assumption (5.26), for which a slightly stronger result is available. This very simple special case also

serves as an outline and motivation for the general case. Also, readers who are only interested in Theorems 6.2 and 1.2 and Corollary 6.4 but not in Theorem 6.6 may read the proof of Lemma 5.6 and skip the calculations of Lemma 5.7.

LEMMA 5.6. *Assume, along with the notation of Theorem* 1.1*, that*

$$\text{rank } C = \text{rank}[B^*, C] = \text{rank}[C, D].$$

Then there exists a matrix Z_0 *satisfying simultaneously the inertia-preserving condition*

$$\pi(P(Z_0)) = \max\{\pi, \pi'\}, \qquad \nu(P(Z_0)) = \max\{\nu, \nu'\}, \tag{5.32}$$

and (using the notation of (5.5) *and* (5.6)*) the two minimal* rank *conditions*

$$\text{rank}(N'(Z_0)) = \text{rank}(R_1), \qquad \text{rank}(N''(Z_0)) = \text{rank}(R_2). \tag{5.33}$$

Such a completion also satisfies the kernel condition

$$\text{Ker } P(Z_0) \supset \text{Ker } (R_1 \oplus I) \ + \ \text{Ker } (I \oplus R_2).$$

Proof. Since (5.26) implies Corollary 5.2, P admits inertia-preserving completions (5.32). The fact that rank(R_1) and Rank(R_2) are the minimal completion ranks for N' and N'' is obvious. To prove that conditions (5.32)–(5.33) are attainable simultaneously, we re-examine the proof of Theorem 1.1, and reduce the situation to Lemma 4.6, where a positive answer is available.

We assume the rank condition (5.26), which implies $\hat{\Delta} = \hat{d} = 0$, and follow the proof of Theorem 1.1. The matrices B'' and D'' in (5.10) turn out to be zero:

$$H = \begin{pmatrix} A & B' & 0 & Z \\ B'^* & C' & 0 & D' \\ 0 & 0 & 0 & 0 \\ Z^* & D'^* & 0 & E \end{pmatrix},$$

of block sizes $(\alpha, \text{rank } C, \delta(C), \gamma)$. Now removing the zero row and column and then taking the Schur complement with respect to C' yields

$$\hat{H}(Y) = \begin{pmatrix} F & Y \\ Y^* & G \end{pmatrix}, \qquad \begin{aligned} F &= A - B'C'^{-1}B'^*, \\ G &= E - D'^*C'^{-1}D', \\ Y &= Z - B'C'^{-1}D'. \end{aligned}$$

We get, therefore,

$$\text{rank } H = \text{rank } \hat{H}(Y) + \text{rank } C, \tag{5.34}$$

$$\begin{aligned} \text{rank } N' &= \text{rank } C + \text{rank } \begin{pmatrix} F & Y \end{pmatrix}, \\ \text{rank } N'' &= \text{rank } C + \text{rank } \begin{pmatrix} Y \\ G \end{pmatrix}. \end{aligned}$$

Lemma 4.8 shows that a matrix $Y = Y_0$ exists for which $\hat{H}(Y)$, $\begin{pmatrix} F & Y_0 \end{pmatrix}$, and $\begin{pmatrix} Y_0 \\ G \end{pmatrix}$ are simultaneously minimum rank completions. Choosing $Z_0 = Y_0 + B'C'^{-1}D'$, we see from (5.34) that Z_0 minimizes all the three ranks involved.

It remains to show the kernel condition. First, we observe, for all Z, that

$$\text{Ker } P(Z) \supset \text{Ker } (N'(Z) \oplus I) \quad \text{and} \quad \text{Ker}(N'(Z)) \subset \text{Ker}(R_1).$$

For Z_0 just obtained, we actually have (5.33), hence the second containment must be equality: $\text{Ker}(N'(Z)) = \text{Ker}(R_1)$. Now the first containment becomes

$$\text{Ker } P(Z_0) \supset \text{Ker } (N'(Z_0) \oplus I) = \text{Ker } (R_1 \oplus I).$$

Similarly Ker $P(Z_0) \supset$ Ker $(I \oplus R_2)$. Thus

$$\text{Ker } P(Z_0) \supset \text{Ker } (R_1 \oplus I) \ + \ \text{Ker } (I \oplus R_2). \qquad \square$$

In the general situation, when the simplifying assumption rank $C = \text{rank}[B^*, C] = \text{rank}[C, D]$ is not assumed, a simultaneous minimal rank solution still exists, but it is not necessarily an inertia-preserving solution, and the additional kernel condition cannot be guaranteed.

LEMMA 5.7 (a simultaneous minimal rank completion lemma). *With the notation of Theorem* 1.1*, the minimal completion* rank*s for* P, N'*, and* N'' *are, respectively,*

$$\underline{r}(P) = p_0 + n_0, \qquad \underline{r}(N') = \text{rank}(R_1) + \Delta', \qquad \underline{r}(N'') = \text{rank}(R_2) + \Delta''.$$

Moreover, there exists a matrix Z_0 *which produces these* rank*s simultaneously.*

Proof. Assume the notation of Theorem 1.1 and Observation 5.1. First we verify the expressions for the minimal ranks involved. Corollary 5.1 implies that $\underline{r}(p) = p_0 + n_0$ for all bordered matrices. Using our definitions of Δ' and Δ'', the identities $\underline{r}(N') = \Delta' + \text{rank}\, R_1$ and $\underline{r}(N'') = \Delta'' + \text{rank}\, R_2$ are obvious.

Having computed the minimum completion ranks for these three matrices, we now demonstrate that the three minimum ranks can be achieved simultaneously. Our plan is to perform all the steps of the proof of Theorem 1.1 simultaneously on the three matrices involved. We call a step *permissible* if a completion exists which preserves the three minimal ranks. As will be seen, not all steps are permissible, and some modification will be necessary.

The reduction of H to $\hat{H}$ in (5.12) is permissible, since C is a common block in all three matrices. Besides $\hat{H}$, this reduction applied to N' and N'' yields

$$\hat{H}_1 = \begin{pmatrix} F_{11} & F_{12} & F_{13} & I_{\Delta''} & 0 & 0 & 0 & X_{11} & X_{12} & X_{13} \\ F_{12}^* & F_{22} & F_{23} & 0 & I_b & 0 & 0 & X_{21} & X_{22} & X_{23} \\ F_{13}^* & F_{23}^* & F_{33} & 0 & 0 & 0 & 0 & X_{31} & X_{32} & X_{33} \\ I_{\Delta''} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_b & 0 & 0 & 0 & 0 & 0 & 0 & I_b & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{\Delta'} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\hat{H}_2 = \begin{pmatrix} I_{\Delta''} & 0 & 0 & 0 & X_{11} & X_{12} & X_{13} \\ 0 & I_b & 0 & 0 & X_{21} & X_{22} & X_{23} \\ 0 & 0 & 0 & 0 & X_{31} & X_{32} & X_{33} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_b & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I_{\Delta'} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & G_{11} & G_{12} & G_{13} \\ 0 & I_b & 0 & 0 & G_{12}^* & G_{22} & G_{23} \\ 0 & 0 & I_{\Delta'} & 0 & G_{13}^* & G_{23}^* & G_{33} \end{pmatrix}.$$

Using (2.3), these operations preserve ranks:

$$\text{rank } N = \text{rank } \hat{H}_1, \qquad \text{rank } N' = \text{rank } \hat{H}_2.$$

Our aim now is to minimize simultaneously rank $\hat{H}$ in (5.12) and rank $\hat{H}_1$ and rank $\hat{H}_2$ above.

The passage from $\hat{H}$ in (5.12) to H' in (5.14) is also permissible, and may be followed by a similar passage from $\hat{H}_i$ to new matrices H'_i, where all the F, G, X entries located in first and last block rows and columns in $\hat{H}_i$ are made zero. We also discard zero rows and columns in these matrices (the seventh block coordinate in H').

Permissibility is violated in the passage from H' to H'' in (5.15). More precisely, complementation of H' with respect to coordinates 1, 4, 6, 10 is permissible; unfortunately, symmetric complementation with respect to coordinates 2 and 5 is not permissible, since these coordinates are not present in both H'_1 and H'_2. Consequently, the proof of Theorem 1.1 has to be modified: we perform on H' non-hermitian complementation (2.3) with respect to block rows 1, 2, 5, 6 and block columns 4, 5, 9, 10, i.e. with respect to the matrix

$$H'_{[1,2,5,6][4,5,9,10]} = \begin{pmatrix} I_{\Delta''} & 0 & 0 & 0 \\ 0 & I_b & X_{22} & 0 \\ 0 & 0 & I_b & 0 \\ 0 & 0 & 0 & I_{\Delta'} \end{pmatrix},$$

where $b = d' - \Delta''$. The Schur complement of H' with respect to $H'_{[1,2,5,6][4,5,9,10]}$ is

$$\begin{pmatrix} 0 & F_{23}^* & F_{33} & 0 & X_{31} \\ I_{\Delta''} & 0 & 0 & 0 & 0 \\ 0 & X_{21}^* & X_{31}^* & 0 & G_{11} \\ 0 & X_{22}^* & X_{32}^* & 0 & G_{12}^* \\ 0 & 0 & 0 & I_{\Delta'} & 0 \end{pmatrix} - \begin{pmatrix} 0 & 0 & X_{32} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & G_{12} & 0 \\ 0 & I_b & G_{22} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \times \begin{pmatrix} I_{\Delta''} & 0 & 0 & 0 \\ 0 & I_b & -X_{22} & 0 \\ 0 & 0 & I_b & 0 \\ 0 & 0 & 0 & I_{\Delta'} \end{pmatrix}$$

$$\times \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & F_{22} & F_{23} & 0 & X_{21} \\ 0 & I_b & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & X_{32} + F_{23}^* & F_{33} & 0 & X_{31} \\ I_{\Delta''} & 0 & 0 & 0 & 0 \\ 0 & G_{12} + X_{21}^* & X_{31}^* & 0 & G_{11} \\ 0 & G_{22} + F_{22} - X_{21}^* + X_{22}^* & X_{32}^* + F_{23} & 0 & X_{21} + G_{12}^* \\ 0 & 0 & 0 & I_{\Delta'} & 0 \end{pmatrix}.$$

This matrix is of the general form

$$\begin{pmatrix} 0 & W_1 & F_{33} & 0 & W_2 \\ I_{\Delta''} & 0 & 0 & 0 & 0 \\ 0 & W_3 & W_2^* & 0 & G_{11} \\ 0 & W_4 & W_5 & 0 & W_3^* \\ 0 & 0 & 0 & I_{\Delta'} & 0 \end{pmatrix},$$

where the W_i 's are unspecified. Indeed, it is easy to see that any arbitrary choice of the W_i's can be achieved by appropriate choice of the X_i's. The respective Schur complements of H'_1 and H'_2 with respect to $H'_{[1,2,5,6][4,5,9,10]}$ turn out to be

$$\tilde{H}_1 = \begin{pmatrix} 0 & W_1 & F_{33} & 0 & W_2 \\ I_{\Delta''} & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \tilde{H}_2 = \begin{pmatrix} 0 & W_2 \\ 0 & 0 \\ 0 & G_{11} \\ 0 & W_3^* \\ I_{\Delta'} & 0 \end{pmatrix}.$$

Taking W_1, W_3, W_4, and W_5 equal to zero is obviously a minimal rank choice for all three matrices. We have reduced the original problem to the simpler problem of simultaneously minimizing the ranks of the three matrices

$$\begin{pmatrix} F_{33} & W_2 \\ W_2^* & G_{11} \end{pmatrix}, \qquad \begin{pmatrix} F_{33} & W_2 \end{pmatrix}, \qquad \begin{pmatrix} W_2 \\ G_{11} \end{pmatrix}.$$

Reduction to Lemma 4.8 is completed. □

In section 6 we shall use the following weakened form of Lemma 5.7, which has better propagation properties.

COROLLARY 5.8 (propagation of internal inequalities of bordered matrices). *Given the notation of Theorem* 1.1 *and Observation* 5.1*, then there exists a matrix* Z_1 *such that*

$$\begin{array}{ll} \text{rank } N'(Z_1) \le \hat{\Delta} + \text{rank } R_1, & \pi(P(Z_1)) \le \hat{\Delta} + \max\{\pi, \pi'\}, \\ \text{rank } N''(Z_1) \le \hat{\Delta} + \text{rank } R_2, & \nu(P(Z_1)) \le \hat{\Delta} + \max\{\nu, \nu'\}. \end{array} \tag{5.35}$$

Corollary 5.8 follows directly from Lemma 5.7 and Observation 5.1. □

5.6. Solvability of the Riccati equation. We end this section with a small contribution connected to the theory of Lyapunov and Riccati equations (see [LR]).

LEMMA 5.9. *Given matrices* A, B, *and* C *of sizes* $\alpha \times \alpha$, $\alpha \times \beta$, *and* $\beta \times \beta$, *respectively, with* A *and* C *hermitian, define*

$$H = \begin{pmatrix} A & B \\ B^* & C \end{pmatrix}, \qquad r = \text{rank} \begin{pmatrix} B \\ C \end{pmatrix}.$$

Then the possible inertias of matrices of the form $P(Z) = A + BZ^* + ZB^* + ZCZ^*$, *with arbitrary* Z, *form the septagon*

$$\begin{array}{ccccccccc} \pi(A) - r & \le & \pi(P(Z)) & \le & \min\{\alpha, \pi(A)\}, & -\nu(A) \le & \pi(Z) - \nu(P(Z)) & \le & \pi(A), \\ \nu(A) - r & \le & \nu(P(Z)) & \le & \min\{\alpha, \nu(A)\}, & & \pi(P(Z)) + \nu(P(Z)) & \le & \alpha. \end{array}$$

Proof. This is an easy corollary of Theorem 1.1, using complementation on the last two block coordinates of the bordered matrix

$$\begin{pmatrix} -A & B & Z \\ B^* & C & I \\ Z^* & I & 0 \end{pmatrix}. \qquad \square$$

In [CG3] the values of $\underline{n}$ and $\overline{n}$ were determined for this case.

COROLLARY 5.10. *The Riccati equation* $A + BZ^* + ZB^* + ZCZ^* = 0$ *is solvable if and only if* $\max\{\pi, \nu\} \le r$.

These results apply also for the Lyapunov or Stein equations: simply assume that C or B is a zero matrix. We emphasize, however, that in the classical context of these equations Z is assumed hermitian (at least), and then it is not clear whether this puts additional restrictions on the set of inertias.

6. Some completion results for general band matrices. In this section, we consider hermitian matrices with general block band or "staircase structure," where again the blocks may vary in size. We follow the method of Dym and Gohberg [DG], which decomposes the completion process into a succession of simple steps, each involving the completion of one bordered submatrix. Combining this procedure with the results of section 5, we are able to draw several interesting conclusions:

(I) In section 6.3 we identify certain classes of staircase hermitian matrices which admit *inertia-preserving completions*; these are completions which inherit their inertia values p and n from (possibly two different) specified blocks of the given partial matrix P. Such completions are obviously minimal rank completions (see Theorem 6.2 and Corollaries 6.3 and 6.8). These results generalize the (scalar) band hermitian completion results of the second author in [D6].

(II) In section 6.3 we consider block band or staircase matrices for which all maximal specified hermitian submatrices are invertible. We show that such matrices admit all the possible completion inertias consistent with Poincaré's inequalities (see Theorem 1.2) which includes inertia-preserving completions. These results generalize the (scalar) band hermitian completion results of the second author in [D6].

(III) Not every partial matrix admits inertia-preserving completions. In section 6.4 we establish modest upper bounds on the minimal possible rank for hermitian completions of staircase hermitian matrices.

6.1. The staircase matrix notation. In dealing with staircase partial matrices, we shall adhere to the following notation and observations, which shall collectively be referred to as the *staircase notation.*

(I) We recognize the bordered matrix of Theorem 1.1 in each pair (R_i and R_{i+1}) of successive maximal hermitian submatrices. We therefore define the s bordered partial submatrices P_i to be

$$P_i = \begin{pmatrix} A_i & B_i & ? \\ B_i^* & C_i & D_i \\ ? & D_i^* & E_i \end{pmatrix},$$

where these specified submatrices $A_i, B_i, C_i, D_i, E_i,\ i = 1, \ldots, s,$ of R are

$$A_i = \{a_{m,l}\}_{m,l=j_{i+1}}^{j_{i+1}-1}, \qquad B_i = \{a_{m,l}\}_{m=j_i,\ l=j_{i+1}}^{j_{i+1}-1,\ k_i}, \qquad C_i = \{a_{m,l}\}_{m,l=j_{i+1}}^{k_i},$$

$$D_i = \{a_{m,l}\}_{m=j_{i+1},\ l=k_i+1}^{k_i,\ k_{i+1}}, \qquad E_i = \{a_{m,l}\}_{m,l=k_i+1}^{k_{i+1}}.$$

As in the notation of Theorem 1.1, we observe that the R_i are the specified block submatrices:

$$R_i = \begin{pmatrix} A_i & B_i \\ B_i^* & C_i \end{pmatrix}, \qquad R_{i+1} = \begin{pmatrix} C_i & D_i \\ D_i^* & E_i \end{pmatrix},$$

and that each C_i is the overlap of R_i and R_{i+1}. The submatrices R_1, C, and R_2 of the notation of Theorem 1.1 correspond to the submatrices R_i, C_i, and R_{i+1}, respectively of each bordered submatrix P_i.

(II) We shall denote by $P_i(Z_i)$ the completions of P_i, using Z_i in the right upper block of P_i. We denote the inertias of $P_i(Z_i)$ by $(\pi_i', \nu_i', \delta_i')$.

(III) In dealing with the ith bordered submatrix P_i, the incremental ranks $\Delta', \Delta'', \hat{\Delta}$, and $d', d'', \hat{d}$ defined in the notation of Theorem 1.1 and Observation 5.1 will be distinguished by the subscript i.

6.2. The diagonal completion formalism. (I) A *diagonal (partial) completion* R_+ of an s-step staircase partial matrix R is an $(s-1)$-step staircase partial hermitian matrix obtained by completing all the s bordered matrices P_i of R. This entails the addition of a matched pair of whole block diagonals alongside the specified band. In a diagonal completion the different bordered completions $P_i(Z_i)$ are independent of each other, that is, the Z_i do not overlap.

(II) A standard procedure for completing a staircase matrix is by a succession of diagonal completions. Let F denote a full hermitian completion of R. We may obtain F via a chain of $s+1$ staircase partial matrices,

$$R, R_+, R_{++}, \dots, F. \tag{6.1}$$

Each matrix in (6.1) is obtained from its precursor via diagonal completion, and its staircase pattern is reduced by one step. We shall distinguish between the relevant submatrices P_i, C_i, R_i of each matrix in this chain by attaching to them the appropriate number of (subscript) plus signs.

(III) The $N+1-st$ matrix ($1 \le N \le s$) in the above chain is called an *N-diagonal (partial) completion* of R.

(IV) We will identify certain submatrices of R with their counterparts in R_+. For example, the completed bordered matrices $P_i(Z_i)$ of R will be identified with the matrices R_{+i} of R_+. In addition, the maximal submatrices R_i of R will be identified with the submatrices C_{+i} of R_+. The ordering of these matrices will always be from top left to bottom right.

This technique of completing a hermitian scalar band matrix by adding successive pairs of diagonals was developed by Dym and Gohberg in [DG]. This technique of completing a hermitian scalar band matrix was later used in many papers, including [D5], [D6], [DG], [EGL1], [EGL1], and [EL]. We shall use the more general staircase or general block band approach of the appendix of [JR2].

In applying the bordered case (for example, Theorem 1.1) to sections of a band matrix or the more general staircase matrices, the key concept is *propagation*. Those properties which survive a single (Theorem 1.1) completion step will by induction survive the full completion process. Our results in this section are all based on properties which propagate.

6.3. Inertia-preserving completions. We call an N-step completion R' of R *inertia preserving* if it satisfies the equations

$$\max_i\{\ \pi(R_i)\} = \max_i\ \{\pi(R_i')\}, \qquad \max_i\{\ \nu(R_i)\} = \max_i\{\ \nu(R_i')\}. \tag{6.2}$$

In particular, if F is a fully specified completion, then it is inertia-preserving if

$$\max\{\ \pi(R_i)\} = \pi(F), \qquad \max\{\ \nu(R_i)\} = \nu(F). \tag{6.3}$$

Such a completion is necessarily a minimal rank completion.

Not every partial staircase matrix admits an inertia-preserving completion. In fact, Example 6.2 will present an infinite sequence of partial staircase matrices whose maximal specified submatrices all have rank 2, but all the full hermitian completions are invertible.

LEMMA 6.1 (propagation of inertia preservation). *Given an s-step hermitian staircase matrix R, using the notation of section* 6.1*, suppose that the blocks of R satisfy these propagation equations:*

$$\text{Rank } (B_i^*, C_i) = \text{Rank } C_i = \text{Rank } (C_i, D_i) \tag{6.4}$$

for each $i = 1, 2, \dots, s$. Then (using the notation of section 6.2*) we have the following:*

(i) *There exists a one step inertia-preserving completion R_+ of R for which*

$$\text{Rank } (B_{+i}^*, C_{+i}) = \text{Rank } C_{+i} = \text{Rank } (C_{+i}, D_{+i}) \tag{6.5}$$

for each $i = 1, 2, \ldots, s-1$.

(ii) *This completion* R_+ *satisfies the kernel condition*

$$\text{Ker } R_{+i} \supset \text{Ker } (R_i \oplus I) \ + \ \text{Ker } (I \oplus R_{i+1}).$$

Proof. The proof of (i) is a straightforward application of Lemma 5.6 repeated s times. Equation (6.4) implies that each P_i fulfills the hypothesis of that lemma. Therefore each P_i admits a completion $P_i(Z_i)$ which achieves simultaneously

$$\text{rank } P = \max\{\pi, \pi'\} + \max\{\nu, \nu'\}, \quad \text{rank } N' = \text{rank } R_1, \quad \text{rank } N'' = \text{rank } R_2,$$

using the notation of (5.5) and (5.6). For P_i, the condition $\text{rank } N'' = \text{rank } R_2$ translates into

$$\text{Rank } C_{+i} = \text{Rank } (C_{+i}, D_{+i}).$$

For P_{i+1}, the condition $\text{rank } N' = \text{rank } R_1$ translates into

$$\text{Rank } (B^*_{+i}, C_{+i}) = \text{Rank } C_{+i}.$$

Together this is precisely (6.5). The condition $\pi(P) = \max\{\pi, \pi'\}$ of Lemma 5.4 applied to each P_i will establish

$$\max_i \{\pi(R_i)\} = \max_i \{\pi(R_{+i})\}.$$

Part (ii) follows from part (i), (3.3), and Lemma 5.6. ∎

We observe that (6.4) is a *propagation* condition: Lemma 6.1 shows that this condition can be made to survive a single diagonal completion. Repeating Lemma 5.9 s times along the chain (6.1), we get the following.

THEOREM 6.2 (inertia-preserving completions). *Given an* s*-step hermitian staircase matrix* R *fulfilling the propagation* (6.4)*,*

$$\text{rank } (B^*_i, C_i) = \text{rank } C_i = \text{rank } (C_i, D_i).$$

Then R *admits an inertia-preserving fully specified completion* F *for which*

$$\underline{p} = \pi(F) = \max\{\pi_i\}, \quad \underline{n} = \nu(F) = \max\{\nu_i\}, \quad \underline{r} = \text{rank } F = \max\{\pi_i\} + \max\{\nu_i\}.$$

Moreover, for this completion F*,* Ker F *contains all the appropriate kernels of the form* Ker $(I \oplus R_i \oplus I)$.

Theorem 6.2 and its proof are largely a block generalization of Dancis' proof in [D6]. The Poincaré inequalities show that the expressions in the theorem are lower bounds for $\underline{p}, \underline{n}, \underline{r}$. Theorem 6.2 shows that they are achieved.

COROLLARY 6.3. *If all the matrices,* C_i *and* R_i *in a staircase matrix* R*, have the same* rank r*, then there exists a hermitian completion* F *with* rank r.

Proof. The condition that all the C_i and R_i matrices have the same rank implies (6.4) and hence Theorem 6.2 is applicable. ∎

This corollary was established for (hermitian and non-hermitian) completions of hermitian and non-hermitian, respectively, band matrices in [EL].

An important special case where Theorem 6.2 is applicable is when all the C_i submatrices of R are invertible.

COROLLARY 6.4 (inertia-preserving completions). *Given an* s*-step staircase hermitian matrix* R*. Suppose that the* s *submatrices* C_i *of* R *are all invertible. Then* R *admits an inertia-preserving completion* F *whose kernel contains all the appropriate kernels of the form* Ker $(I \oplus R_i \oplus I)$.

Proof. Since all the C_i are invertible, the propagation (6.4) holds and Theorem 6.2 applies. ∎

6.4. Incremental bounds on inertia growth. In general, even assuming a minimal rank completion in each step, the ranks of the matrices in (6.1) may increase. At present, we cannot compute the minimal rank for the completions of a general staircase matrix or even for a general scalar band matrix. The reason is lack of propagation: the inertias of $P_i(Z_i)$ do not depend exclusively on the inertias of R_i and C_i, as is evident from Theorem 1.1. However, use of Corollary 5.8 will enable us to obtain an upper bound on the inertia increase.

In the next observation and lemma, the B_{+i}, C_{+i}, and D_{+i} matrices are the B_i, C_i, and D_i matrices of a diagonal completion R_+; the d_{+i} will be

$$d'_{+i} = \text{rank } [B^*_{+i}, C_{+i}] - \text{rank } C_{+i}, \quad d''_{+i} = \text{rank } [C_{+i}, D_{+i}] - \text{rank } C_{+i},$$

$$\text{and} \quad \hat{d}_{+i} = \max\{d'_{+i}, d''_{+i}\},$$

which is consistent with our general notation.

Observation 6.1. Let R be a staircase matrix, together with the notation of sections 6.1 and 6.2. Suppose that R_+ is a diagonal completion of R for which all bordered completions $P_i(Z_i)$ satisfy the simultaneous minimal rank completion Lemma 5.7. Then the incremental ranks of the bordered submatrices of R and of R_+ are related via

$$d'_{+i} = \Delta''_i \qquad \text{and} \qquad d''_{+i-1} = \Delta'_i. \tag{6.6}$$

Proof. We have the following connections between R-related and R_+-related objects:

$$B_{+i} = \begin{pmatrix} B_i & Z_i \end{pmatrix}, \qquad C_{+i} = \begin{pmatrix} C_i & D_i \\ D_i^* & E_i \end{pmatrix} = R_{i+1},$$
$$N''_i(Z_i) = \begin{pmatrix} B_{+i} & C_{+i} \end{pmatrix}^*, \quad d'_{+i} = \text{rank } [B^*_{+i}, C_{+i}] - \text{rank } C_{+i}.$$

From Lemma 5.7 we note that

$$\text{rank } N''_i(Z_i) = \text{rank } R_{i+1} + \Delta''_i.$$

Combining these equations yields $d'_{+i} = \Delta''_i$. The other equation may be similarly observed. ∎

LEMMA 6.5 (propagation of incremental ranks). *Let R be an s-step hermitian staircase matrix. We use the notation*

$$\hat{d}_i = \max\{\text{rank}[C_i, D_i],\ \text{rank}[B_i^*, C_i]\} - \text{rank}C_i,$$

$$\hat{d}_{+i} = \max\{\text{rank}[C_{+i}, D_{+i}], \text{rank}[B^*_{+i}, C_{+i}]\} - \text{rank}C_{+i},$$

which is consistent with our bordered and band matrix notation. Set $\hat{d} = \max\{\hat{d}_i\}$, $\hat{d}_+ = \max\{\hat{d}_{+i}\}$. Then there exists a diagonal completion R_+ of R for which

$$\hat{d} + \max\{\pi_i\} \geq \max\{\pi_{+i}\}, \qquad \hat{d} + \max\{\nu_i\} \geq \max\{\nu_{+i}\}, \qquad \hat{d} \geq \hat{d}_+. \tag{6.7}$$

Proof. The proof is a straightforward application of Corollary 5.8. The definition of $\hat{d}$ and (6.6) implies that $\hat{d} \geq \hat{\Delta}_i$. Therefore for each P_i there exists a completion $P_i(Z_i)$ which achieves simultaneously

$$\pi(P(Z)) \leq \hat{d} + \max\{\pi, \pi'\}, \qquad \nu(P(Z)) \leq \hat{d} + \max\{\nu, \nu'\} \tag{6.8}$$

and

$$\text{rank } N' \leq \hat{d} + \text{rank } R_1, \quad \text{rank } N'' \leq \hat{d} + \text{rank } R_2. \tag{6.9}$$

The last inequalities (6.9) translate to $d'_{+i} \leq \hat{d}$ and $d''_{+i} \leq \hat{d}$, which repeated over all i implies that $\hat{d}_+ \leq \hat{d}$. The former inequalities (6.8) prove the rest of (6.7). ∎

We observe that inequalities (6.7) combine to form a *propagation* condition: Lemma 6.5 shows that these inequalities may be transferred from a staircase matrix to a diagonal completion. Repeating Lemma 6.5 until R is fully completed, we get the following.

THEOREM 6.6 (incremental bounds on inertia growth). *Given an s-step hermitian staircase matrix R (together with the notation of section* 6.1*), then there exists a hermitian completion F of R whose inertia (p, n, d) satisfies*

$$p \leq s \max\{\hat{d}_i\} + \max\{\pi(R_i)\}, \qquad n \leq s \max\{\hat{d}_i\} + \max\{\nu(R_i)\}.$$

COROLLARY 6.7. *Given a hermitian staircase matrix R (together with the notation of section* 6.1*), suppose there is an integer t such that*

$$\text{rank } R_i \leq 2t + 1 + \text{rank } C_i \quad \textit{and} \quad \text{rank } R_{i+1} \leq 2t + 1 + \text{rank } C_i, \quad i = 1, 2, \ldots, s,$$

then there exists a hermitian completion F of R such that

$$\pi(F) \leq st + \max\{\pi(R_i)\}, \quad \textit{and} \quad \nu(F) \leq st + \max\{\nu(R_i)\}.$$

Proof. The hypotheses and (5.3) provide rank $C_i + 2\hat{d}_i \leq$ rank $R_i \leq 2t + 1 +$ rank C_i, hence $2\hat{d}_i \leq 2t+1$. Since both t and $\hat{d}_i$ are *integers*, this inequality becomes $\hat{d}_i \leq t$. In this way, one sees that $t \geq \max\{\hat{d}_i\}$ and the theorem is applicable.

The case $t = 0$ is of particular interest.

COROLLARY 6.8. *Given a hermitian staircase matrix R (together with the notation of section* 6.1*), suppose that*

$$\text{rank } R_i \leq 1 + \text{rank } C_i \quad \textit{and} \quad \text{rank } R_{i+1} \leq 1 + \text{rank } C_i, \quad i = 1, 2, \ldots, s;$$

then there exists an inertia-preserving hermitian completion F of R .

The fact that Theorem 6.6 is best possible without additional hypotheses is demonstrated by the next example.

EXAMPLE 6.2. *Consider the matrix $R(U) = \left(\begin{smallmatrix} 0 & I+U \\ I+U^* & 0 \end{smallmatrix}\right)$, where U is an $(s+1) \times (s+1)$ strictly upper triangular matrix. We consider $R(U)$ as a partial s-step band matrix, in which U is unspecified. In the notation of Theorem* 6.6*, all $\hat{d}_i = 1 = \pi(R_i) = \nu(R_i)$. Thus, by this theorem, we expect to find a completion with*

$$p \leq s + 1, \quad n \leq s + 1. \tag{6.10}$$

However, since det $R(U) = \pm 1$*, we have $p + n = 2s + 2$, hence* (6.10) *can only be satisfied with equality. In fact, using the extended Poincaré inequalities, we see that every completion $R(U)$ satisfies* (6.10) *with equality.*

6.5. Proof of Theorem 1.2: A staircase of invertible maximal hermitian submatrices. Throughout this section we shall assume that R is an s-step staircase partial matrix with all maximal submatrices R_i invertible. In this case, all inertias compatible with Poincarés inequalities are achievable with a hermitian completion.

First we present the minimal-rank inertia-preserving case as the next lemma.

LEMMA 6.9 (an inertia-preserving lemma). *Given an s-step staircase hermitian matrix R (together with the notation of section* 6.1*), suppose that each of the maximal submatrices $R_1, R_2, \ldots, R_s$ of R is an invertible matrix. Then there is a hermitian completion F of R such that*

$$\pi(F) = \text{Max}\{\pi(R_i),\ i = 1, 2, \ldots, s\} \quad \textit{and} \quad \nu(F) = \text{Max}\{\nu(R_i),\ i = 1, 2, \ldots, s\}$$

and such that

$$\text{Ker } F \quad \textit{contains} \quad \text{Ker } R_1 + \text{Ker } R_2 + \ldots + \text{Ker } R_s.$$

Proof. We use Corollary 5.10 s times as we construct an inertia-preserving diagonal completion R_+ of R. Then R_+ will satisfy the hypotheses of Corollary 6.4, which will produce the desired hermitian completion F with $\pi(F) = \pi(F) = \text{Max}\{\pi(R_i)\}$ and $\nu(F) = \text{Max}\{\nu(R_i)\}$. □

Proof of Theorem 1.2. We will use Corollary 5.5 repeatedly to construct successive diagonal completions with invertible maximal matrices and increasing inertias.

Construction of the (first) diagonal completion (R_+).

Case 1. If the size of $P_i(Z_i) < p+r$, then Corollary 5.5 is used to choose a matrix Z_i such that $P_i(Z_i)$ is an invertible matrix with

$$\pi(P_i(Z_i)) \le p, \quad \nu(P_i(Z_i)) \le n,$$

$$\pi(P_i(Z_i)) \ge \max\{\pi(R_i), \pi(R_{i+1})\}, \quad \nu(P_i(Z_i)) \ge \max\{\nu(R_i), \nu(R_{i+1})\}.$$

Case 2. If the size of $P_i(Z_i) \ge p+r$, then Corollary 5.5 is used to choose a matrix Z_i such that

$$\text{In } (P_i(Z_i)) = (p, n, *).$$

Depending only on its size, $P_i(Z_i)$ may be an invertible or a noninvertible matrix.

In both cases, the new $\{C_{+i}\}$ are the previous maximal submatrices $\{R_i\}$, and hence all the new $\{C_{+i}\}$ of R_+ are invertible matrices.

If Case 2 occurred at least once, then the desired positivity and negativity has been achieved. Then Corollary 6.4 applied to R_+ will provide the desired full completion F, with $\text{In } F = (p, n, *)$.

If no Case 2 has occurred, only Case 1, then all the maximal submatrices of R_+ are invertible.

In this manner, one constructs a number of successive diagonal completions *until* Case 2 is used. With each successive diagonal completion, the values of the positivity and negativity grow. At some point, at least one of the new $\pi(R_i)$ and one of the new $\nu(R_j)$ (for the latest successive diagonal completion $R_+ \ldots_+$) will reach the desired values p and n. Then $\pi(R_+ \ldots_+) = p$ and $\nu(R_+ \ldots_+) = n$. This will occur when Case 2 is used, possibly sooner. With the possible exception of the current diagonal completion, all the maximal specified submatrices of the various successive diagonal completions were invertible (since only Case 1 was used). Therefore all the C_i of the current diagonal completion were the *invertible* maximal submatrices of the previous diagonal completion. Therefore Corollary 6.4 is applicable and it completes the proof.

REFERENCES

[AHMR] J. Agler, J. W. Helton, S. McCullough, and L. Rodman, *Positive definite matrices with a given sparsity pattern*, Linear Algebra Appl., 107 (1988), pp. 101–149.

[BGRS] J. A. Ball, I. Gohberg, L. Rodman, and T. Shalom, *On the eigenvalues of matrices with given upper triangular part*, Integral Equations Operator Theory, 13 (1990), pp. 488–497.

[BJL] W. Barrett, C. R. Johnson, and M. Lundquist, *Determinantal formulae for matrix completions associated with chordal graphs*, Linear Algebra Appl., 121 (1989), pp. 265–289.

[C] B. Cain, *The inertia of a Hermitian matrix having prescribed diagonal blocks*, Linear Algebra Appl., 37 (1981), pp. 173–180.

[CD] N. Cohen and J. Dancis, *Maximal rank Hermitian completions of partially specified Hermitian matrices*, Linear Algebra Appl., 244 (1996), pp. 265–276.

[CG1] T. Constantinescu and A. Gheondea, *Minimal signature in lifting of operators* I, J. Operator Theory, 22 (1989), pp. 345–367.

[CG2] T. Constantinescu and A. Gheondea, *Minimal signature in lifting of operators* II, Operator Theory, to appear.

[CG3] T. Constantinescu and A. Gheondea, *The negative signature of some Hermitian matrices*, Linear Algebra Appl., 178 (1993), pp. 17–42.

[CJRW] N. Cohen, C. R. Johnson, L. Rodman, and H. J. Woerdeman, *ranks of completions of partial matrices*, Operator Theory Adv. Appl., 40 (1989), pp. 165–185.

[CS] B. E. Cain, and E. Marques de Sá, *The inertia of a Hermitian matrix having prescribed complementary principal submatrices*, Linear Algebra Appl., 37 (1981), pp. 161–171.

[D1] J. Dancis, *The possible inertias for a Hermitian matrix and its principal submatrices*, Linear Algebra Appl., 85 (1987), pp. 121–151.

[D2] J. Dancis, *On the inertias of symmetric matrices and bounded self-adjoint operators*, Linear Algebra Appl., 105 (1988), pp. 67–75.

[D3] J. Dancis, *Bordered matrices*, Linear Algebra Appl., 128 (1990), pp. 117–132.

[D4] J. Dancis, *Several consequences of an inertia theorem*, Linear Algebra Appl., 136 (1990), pp. 43–61.

[D5] J. Dancis, *Positive semidefinite completions of partial Hermitian matrices*, Linear Algebra Appl., 175 (1992), pp. 97–114.

[D6] J. Dancis, *Choosing the inertias for completions of certain partially specified matrices*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 813–829.

[D7] J. Dancis, *Ranks and Inertias of Hermitian Toeplitz Matrices*, Technical Report TR94-08, Dept. of Math., Univ. of Maryland, 1994.

[DG] H. Dym and I. Gohberg, *Extensions of band matrices with band inverses*, Linear Algebra Appl., 36 (1981), pp. 1–24.

[EGL1] R. Ellis, I. Gohberg, and D. C. Lay, *Band extensions, maximum entropy and the permanence principle*, in Maximum Entropy and Bayesian Methods in Applied Statistics, J. Justice, ed., Cambridge University Press, London, 1986, pp. 131–155.

[EGL2] R. Ellis, I. Gohberg, and D. C. Lay, *Invertible self adjoint extensions of band matrices and their entropy*, SIAM J. Alg. Disc. Methods, 8 (1987), pp. 483–500.

[EL] R. Ellis and D. C. Lay, *rank-preserving extensions of band matrices*, Linear and Multilinear Algebra, 26 (1990), pp. 147–179.

[G] A. Gheondea, *One-step completions of Hermitian partial matrices with minimal negative signature*, Linear Algebra Appl., 173 (1992), pp. 99–114.

[GJSW] B. Grone, C. R. Johnson, E. Marques de Sá, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

[H] E. V. Haynesworth, *Determination of the inertia of a partitioned Hermitian matrix*, Linear Algebra Appl., 1 (1968), pp. 73–81.

[HO] E. V. Haynesworth and A. M. Ostrowski, *On the inertia of some classes of partitioned matrices*, Linear Algebra Appl., 1 (1968), pp. 299–316.

[HJ] R. A. Horn and C. R. Johnson, *Matrix Analysis* I, Cambridge University Press, London, 1985.

[I] I. S. Iohvidov, *Hankel and Toeplitz Matrices and Forms*, Birkhäuser, Boston 1982.

[JR1] C. R. Johnson and L. Rodman, *Inertia possibilities for completions of partial Hermitian matrices*, Linear and Multilinear Algebra, 16 (1984), pp. 179–195.

[JR2] C. R. Johnson and L. Rodman, *Chordal inheritance principles and positive definite completions of partial matrices over function rings*, in Contributions to Operator Theory

and its Appl., I. Gohberg et al., eds., Birkhäuser, Basel, Switzerland, 1988.

[L] M. E. Lundquist, *Zero Patterns, Chordal Graphs and Matrix Completions*, Dissertation, Clemson University, Clemson, SC, 1990.

[LR] P. Lancaster and P. Rosza, *On the matrix equation* $AX + X^*A^* = C$, SIAM J. Alg. Disc. Methods, 4 (1983), pp. 432–436.

[S] E. Marques de Sá, *On the inertia of sums of Hermitian matrices*, Linear Algebra Appl., 37 (1981), pp. 143–159.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 613–639, July 1998

002

CONE INCLUSION NUMBERS*

GEIR NÆVDAL† AND HUGO J. WOERDEMAN‡

Abstract. The introduction of cone inclusion numbers allows one to view seemingly different problems from one general perspective. Using this perspective several new results are obtained, such as: (a) the distance constant for $\mathcal{T}_n \otimes \mathcal{T}_n$, where $\mathcal{T}_n$ denotes the algebra of $n \times n$ strictly upper triangular matrices, is bounded above by $\lceil \log_2 n \rceil + 1$; (b) for every natural number n there exists an $n \times n$ partial correlation matrix for which the largest possible minimal eigenvalue of a completion is $1 - \sqrt{\lfloor \frac{n}{2} \rfloor}$; and (c) the lowest possible entry-wise supremum norm among all $n \times n$ matrices that induce a norm one Schur map is $\frac{1}{\sqrt{n}}$.

Key words. cones of matrices, matrix completion, positive semidefinite matrices, distance constants, sparsity patterns

AMS subject classifications. 15A48, 15A57, 05C50, 47D25, 47D20, 15A60

PII. S0895479896296908

1. Introduction.

An inclusion number may, in its full generality, be introduced as follows: given a cone $\mathcal{C}$ in a vector space V, a set $S \subseteq V$ and a point $a \in V$, define

$$\alpha = \inf\{\, \lambda \in \mathbb{R} : S + \lambda a \subseteq \mathcal{C} \,\}. \tag{1}$$

It is our goal to bring some seemingly different problems together under the notion of cone inclusion number with closely related choices for V, $\mathcal{C}$, a, and S. In fact, we will restrict our attention to the following setting. The Hilbert space V consists of Hermitian matrices with a certain sparsity pattern, $\mathcal{C}$ is a cone derived from the cone of positive semidefinite matrices, a is the identity matrix, and S is the intersection of a second cone $\widehat{\mathcal{C}}$ with the affine space consisting of matrices with 1's on the main diagonal. The cone inclusion numbers that we study measure in some sense how the two cones $\mathcal{C}$ and $\widehat{\mathcal{C}}$ compare in size. (See Figure 1.)

Our starting point was the theory of distance constants and, in particular, the still unresolved question posed by K. R. Davidson [13]: Is there a uniform bound on the distance constants for $\mathcal{T}_n \otimes \mathcal{T}_n$, $n \in \mathbb{N}$? Here $\mathcal{T}_n$ denotes the algebra of strictly upper triangular matrices. It was not a big step to consider Hermitian variations of these types of problems. Let us elaborate on a question that has been posed by C. R. Johnson as a natural follow-up on the results in [19].

It is known that a partially defined matrix may fail to have a positive semidefinite completion even though all of its specified principal submatrices are positive semidefinite. Recall that a *partially defined matrix* (or *partial matrix*) is a matrix with some of its entries specified (real or complex, in our case) numbers and the remaining entries specified free variables (over $\mathbb{R}$ or $\mathbb{C}$). A *completion* of a partially defined matrix is obtained by choosing (real or complex) numbers for the free variables, resulting in an

*Received by the editors January 2, 1996; accepted for publication (in revised form) by V. Mehrmann May 26, 1997; published electronically March 18, 1998.

http://www.siam.org/journals/simax/19-3/29690.html

†Department of Engineering, Stord/Haugesund College, Skåregt. 103, N-5500 Haugesund, Norway. Current address: RF-Rogaland Research, Thormöhlensgt. 55, N–5008 Bergen, Norway (geir.naevdal@rf.no).

‡Department of Mathematics, The College of William and Mary, Williamsburg, Virginia 23187-8795 (hugo@math.wm.edu). The research of this author was supported in part by NSF grant DMS 9500924.

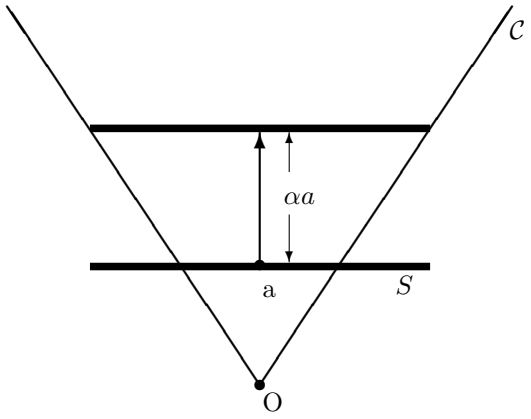


FIG. 1. *The cone inclusion number.*

ordinary matrix. An example of a partially defined matrix with positive semidefinite specified principal submatrices, but without a positive semidefinite, completion is

$$\begin{pmatrix} 1 & 1 & ? & -1 \\ 1 & 1 & 1 & ? \\ ? & 1 & 1 & 1 \\ -1 & ? & 1 & 1 \end{pmatrix}. \tag{2}$$

Indeed, all the specified principal submatrices are 2×2 matrices with 1's on the diagonal, ± 1 off the main diagonal, and are thus all positive semidefinite. However, it is easy to see (see [19]) that (2) does not have a positive semidefinite completion.

It is natural to ask how "bad" the situation can be; more specifically, let A be an $n \times n$ *partial correlation matrix*, i.e., a partial matrix with the property that all its specified principal submatrices are correlation matrices. A *correlation matrix* is a matrix which is positive semidefinite, and with 1's on the diagonal. Consider

$$\nu(A) := \max_{B=B^* \text{completion of } A} \lambda_{\min}(B),$$

where $\lambda_{\min}(B)$ denotes the smallest eigenvalue of the Hermitian matrix B. We study the minimum value that $\nu(A)$ can attain as A varies over the $n \times n$ partial correlation matrices. As we shall see, the lowest possible $\nu(A)$ appears as a cone inclusion number.

We shall provide the following answer to this question. For $r \in \mathbb{R}$, let $\lfloor r \rfloor$ ($\lceil r \rceil$) denote the largest (smallest) integer smaller (greater) or equal to r.

THEOREM 1.1. *For every $n \in \mathbb{N}$ there exists an $n \times n$ partial complex valued correlation matrix A so that*

$$\max_{B=B^* \text{ completion of } A} \lambda_{\min}(B) = 1 - \sqrt{\lfloor \frac{n}{2} \rfloor}. \tag{3}$$

When $n = 2^m$ or $2^m + 1$ for some $m \in \mathbb{N}$, the partial matrix may be chosen to be real.

The partial matrix in (2) provides an illustration of the statement of Theorem 1.1 for the case $n = 4$: the completion of (2) with largest minimal eigenvalue is obtained by putting $? = 0$, and this largest minimal eigenvalue equals $1 - \sqrt{2}$ (this computation may be found in [28]).

We will also prove a partial converse.

THEOREM 1.2. *If A is an $n \times n$ partial correlation matrix A such that all specified submatrices of A have all their eigenvalues $\geq 1 - \frac{1}{n-1}$, then by choosing the unknown entries to be zero one obtains a positive semidefinite matrix A_c which is the sum of rank one positive semidefinite matrices with the same sparsity pattern as A_c. In particular, A has a positive semidefinite completion.*

As a corollary of Theorems 1.1 and 1.2 we may state that

$$2 - n \leq \min \nu(A) \leq 1 - \sqrt{\lfloor \frac{n}{2} \rfloor}, \tag{4}$$

where the minimum is taken over all $n \times n$ complex partial correlation matrices A. In our examples we have not encountered a partial correlation matrix A such that

$$\nu(A) < 1 - \sqrt{\lceil \frac{n}{2} \rceil},$$

and in fact, we conjecture that this cannot occur (for the exact statement, see Conjecture 4.10). In other words, we conjecture that when a partial correlation matrix has all specified submatrices $\geq (1 - \sqrt{1/\lfloor n/2 \rfloor})I$, then A has a positive semidefinite completion. Here I denotes the identity matrix.

There are many other specific instances of cone inclusion numbers, which all represent in some sense a worst case scenario. They include questions such as:

1. What is the lowest minimal eigenvalue among the spectra of graphs with n vertices? (See Theorem 3.1.)

2. What is the largest distance constant among partial matrices of size n? (An upper bound is given in Corollary 5.3.)

3. How far off (in the sense of adding a multiple of the identity matrix) can an $n \times n$ correlation matrix with a given sparsity pattern be from a sum of positive semidefinite rank 1's with the same sparsity pattern? (See Corollary 2.2.)

4. What is the lowest possible entry-wise supremum norm among all $n \times n$ matrices that induce a norm one Schur map? (See Corollary 4.3.)

Our paper is organized as follows. In section 2 we introduce the cone inclusion numbers we study and obtain some of their basic properties. These include a duality result showing, for instance, that question 3 above is equivalent to finding min $\nu(A)$, where A ranges over all $n \times n$ partial correlation matrices (see Corollary 2.2). In section 3 we address the question of finding extreme values for some of the cone inclusion numbers and thereby obtain Theorem 1.2. In section 4 we focus on the cone inclusion numbers most pertinent to the positive semidefinite completion problem and so prove Theorem 1.1. In sections 5 and 6, we turn to the contractive and Toeplitz case, respectively.

2. Cone inclusion numbers. The notions we introduce involve cones in matrix space. The book [6] may be used as a general reference on such cones. All notions defined below may be interpreted as notions in real or complex matrix space (i.e., in $\mathbb{R}^{n \times n}$ or $\mathbb{C}^{n \times n}$). In this section the statements are the same in either setting, making it unnecessary to distinguish between the two. We will therefore suppress any reference to the underlying field. In subsequent sections we will, however, occasionally distinguish between the two cases. We will indicate this by attaching to the notations a subscript or superscript $\mathbb{R}$ or $\mathbb{C}$. For instance, $\mathcal{H}_n^{\mathbb{C}}$, $q_{\mathbb{R}}(P)$, etc. Note that a Hermitian matrix in $\mathbb{R}^{n \times n}$ is just a symmetric matrix.

The *sparsity pattern* (location of required zeroes) of an $n \times n$ Hermitian matrix can be indicated in two ways. The first is via an undirected graph with vertices $\{1, ..., n\}$ which has an edge between vertices i and j ($i \neq j$) precisely when the entries (i, j) and (j, i) are not required to be zero. We will never require the diagonal entries to be zero; however, the graph does not have edges $\{i, i\}$, $i \in \{1, ..., n\}$. The second way is via a subset P of $\{1, ..., n\} \times \{1, ..., n\}$, where (i, j) and (j, i) belong to P if and only if the entries (i, j) and (j, i) are not required to be zero. The pairs $(i, i), i \in \{1, ..., n\}$ always belong to P. Whether or not we represent the sparsity pattern via a graph or via a subset of $\{1, ..., n\} \times \{1, ..., n\}$, we will refer to it as the *pattern* of the matrix/set of matrices. We will use the graph and the subset of $\{1, ..., n\} \times \{1, ..., n\}$ to represent the pattern interchangeably.

A *clique* K of a pattern P is a subset of the form $K = I \times I$ of P. In graph terms a clique corresponds to a full subgraph induced by a subset of the vertices. We say that the clique K of P is *maximal* in P when $K \subseteq L \subseteq P$ with L a clique implies that $K = L$. When A is an $n \times n$ matrix $A = (A_{ij})_{i,j=1}^{n}$ and $K = I \times I$, then $A|K$ denotes the $|I| \times |I|$ principal submatrix

$$A|K = (A_{ij})_{i,j \in I} = (A_{ij})_{(i,j) \in K}.$$

Let $\mathcal{H}_n$ denote the Hilbert space over $\mathbb{R}$ consisting of $n \times n$ (real or complex; see discussion above) Hermitian matrices with inner product

$$\langle A, B \rangle = \operatorname{trace}(AB),$$

where trace M denotes the trace of the square matrix M. For a pattern P we introduce the subspace

$$\mathcal{H}_P = \{\, H \in \mathcal{H}_n : H_{ij} = 0 \text{ for } (i, j) \notin P \,\}.$$

Given a convex cone $\mathcal{C}$ in a Hilbert space $\mathcal{H}$ (i.e., $\mathcal{C} + \mathcal{C} \subseteq \mathcal{C}$ and $\lambda \mathcal{C} \subseteq \mathcal{C}$ for $\lambda \geq 0$) and a subspace W, there are at least six cones in W which one can associate with $\mathcal{C}$, namely, (i) the intersection of $\mathcal{C}$ and W; (ii) the orthogonal projection of $\mathcal{C}$ onto W; (iii) the cone generated by the extreme rays of $\mathcal{C}$ that lie in W; and (iv), (v), and (vi) the duals of (i), (ii), and (iii) in W. Recall that the *dual* of a cone $\tilde{\mathcal{C}}$ in a Hilbert space $\tilde{\mathcal{H}}$ is given by $\tilde{\mathcal{C}}^* = \{\, D \in \tilde{\mathcal{H}} : \langle C, D \rangle \geq 0 \text{ for all } C \in \tilde{\mathcal{C}} \,\}$. We will study the case where $\mathcal{C} = PSD := \{\, H \in \mathcal{H}_n : H \geq 0 \,\}$ consists of the $n \times n$ positive semidefinite matrices, and where $W = \mathcal{H}_P \subseteq \mathcal{H}_n$. Since $PSD^* = PSD$ (i.e., PSD is *self-dual*) the procedure above yields at most four different cones in $\mathcal{H}_P$ as follows.

For a pattern P we define the following cones:

$$\mathcal{A}_P = \{\, A \in \mathcal{H}_P : A|K \geq 0 \text{ for all cliques } K \subseteq P \,\},$$
$$\mathcal{X}_P = \{\, X \in \mathcal{H}_P : X \geq 0 \,\},$$
$$\mathcal{X}_P^* = \{\, Y \in \mathcal{H}_P : \text{ there is a } W \in \mathcal{H}_n \ominus \mathcal{H}_P \text{ such that } Y + W \geq 0 \,\},$$
$$\mathcal{A}_P^* = \left\{ B \in \mathcal{H}_P : B = \sum_{i=1}^{n_B} B_i \text{ where } \; B_i \in \mathcal{X}_P \text{ and } B_i \text{ has rank 1 for all } i \right\}.$$

All four cones are closed. For the first three this follows trivially. In order to show that $\mathcal{A}_P^*$ is closed, one may use an argument similar to the one in the proof of [36, Lemma 1.3].

Note that $\mathcal{H}_n \ominus \mathcal{H}_P = \mathcal{H}_P^{\perp} = \{\, H \in \mathcal{H}_n : H_{ij} = 0 \text{ for } (i, j) \in P \,\}$, so that members of $\mathcal{X}_P^*$ may be viewed as (partial) matrices which have a positive semidefinite

completion. The cones $\mathcal{X}_P$ and $\mathcal{X}_P^*$ are one another's dual in $\mathcal{H}_P$, the proof of which essentially appears independently in [33] and [34]. It may also be viewed as an instance of the general rule that for a closed convex cone $\mathcal{C}$,

$$(\mathcal{C} \cap W)^* = P_W(\mathcal{C}^*),$$

where P_W is the orthogonal projection on the subspace W in $\mathcal{H}$, $*$ in the left-hand side denotes the dual in W, and $*$ in the right-hand side denotes the dual in $\mathcal{H}$. The fact that $\mathcal{A}_P$ and $\mathcal{A}_P^*$ are one another's dual in $\mathcal{H}_P$ is not hard to prove.

In general we have the following relationships:

$$\mathcal{A}_P^* \subseteq \mathcal{X}_P \subseteq \mathcal{X}_P^* \subseteq \mathcal{A}_P.$$

Equality between $\mathcal{X}_P$ and $\mathcal{X}_P^*$ holds only in case the maximal cliques of P are disjoint, and in that case all four cones are equal. Equality between $\mathcal{A}_P^*$ and $\mathcal{X}_P$ (or equivalently $\mathcal{X}_P^*$ and $\mathcal{A}_P$) holds if and only if P is chordal. Recall that an undirected graph P is called *chordal* when every cycle

$$\{i_1, i_2\}, \{i_2, i_3\}, ..., \{i_{m-1}, i_m\}, \{i_m, i_1\}$$

(with $i_k \neq i_l$ for $k \neq l$) with length $m \geq 4$ consisting of edges in P has a chord, i.e., there is an edge $\{i_p, i_q\} (\neq \{i_1, i_m\})$ with $q > p + 1$ in P. The equivalence of $\mathcal{A}_P = \mathcal{X}_P^*$ and P being chordal was obtained in [19]. The equivalent statement that $\mathcal{A}_P^* = \mathcal{X}_P$ if and only if P is chordal was subsequently stated as such in [34]. In order to go beyond the chordal case, the extreme rays of $\mathcal{X}_P$ were studied in [1], [22], [21], and [37].

We have now laid the groundwork for introducing our *cone inclusion numbers*. For any pair of cones $\mathcal{C}_1$ and $\mathcal{C}_2$ from $\mathcal{A}_P^*, \mathcal{X}_P, \mathcal{X}_P^*$, and $\mathcal{A}_P$ we define

$$\alpha(\mathcal{C}_1, \mathcal{C}_2) = \min\{\, \lambda : \mathcal{C}_1\mathcal{D} + \lambda I \subseteq \mathcal{C}_2 \,\}, \tag{5}$$

where for any set $\mathcal{C} \subseteq \mathcal{H}_n$ we denote

$$\mathcal{C}\mathcal{D} = \{\, A \in \mathcal{C} : A_{ii} = 1, \quad i = 1, ..., n \,\}.$$

Figure 2 illustrates the notion.

It is not hard to see that the number $\alpha(\mathcal{C}_1, \mathcal{C}_2)$ is well defined since $\mathcal{C}_2$ is closed, I belongs to its interior (relative to $\mathcal{H}_P$), and $\mathcal{C}_1\mathcal{D}$ is compact. Furthermore, note that the numbers $\alpha(\mathcal{C}_1, \mathcal{C}_2)$ are only of interest if $\mathcal{C}_1 \not\subseteq \mathcal{C}_2$. Clearly, when $\mathcal{C}_1 \subseteq \mathcal{C}_2$ then $\alpha(\mathcal{C}_1, \mathcal{C}_2) \leq 0$. But strict inequality cannot occur, since the matrix E consisting of all 0's, except for 1's on the main diagonal and in one symmetrically located pair of entries, can easily be seen to belong to $\mathcal{A}_P^*$ and to have the property that $E - \epsilon I \notin \mathcal{A}_P$ for any $\epsilon > 0$. So, $\alpha(\mathcal{A}_P^*, \mathcal{A}_P) = 0$. But then $\alpha(\mathcal{C}_1, \mathcal{C}_2) = 0$ follows also for the other possibilities of $\mathcal{C}_1 \subseteq \mathcal{C}_2$. Thus, in all cases $\alpha(\mathcal{C}_1, \mathcal{C}_2) \geq 0$.

We next present different characterizations for $\alpha(\mathcal{C}_1, \mathcal{C}_2)$. For any set $\mathcal{C} \subseteq \mathcal{H}_n$ we denote

$$\mathcal{C}\mathcal{N} = \{\, A \in \mathcal{C} : \operatorname{trace} A = 1 \,\},$$

and $\partial\mathcal{C}$ denotes the topological boundary of $\mathcal{C}$. Furthermore, to relate the result to convex optimization terminology, let $\delta^*(x|C) = \sup\{\, \langle x, c \rangle : c \in C \,\}$ denote the *support function* of a convex set C, and let $\delta(x|C) = \{\begin{smallmatrix} 0 & x \in C \\ \infty & x \notin C \end{smallmatrix}$ denote the *indicator function* (see [35]). In the next proposition we exclude the trivial case $P = \{\, (i, i) : i = 1, ..., n \,\}$.

PROPOSITION 2.1. *Consider* $\mathcal{C}_1, \mathcal{C}_2 \in \{\, \mathcal{A}_P^*, \mathcal{X}_P, \mathcal{X}_P^*, \mathcal{A}_P \,\}$. *Then* $\alpha(\mathcal{C}_1, \mathcal{C}_2)$ *is equal to*

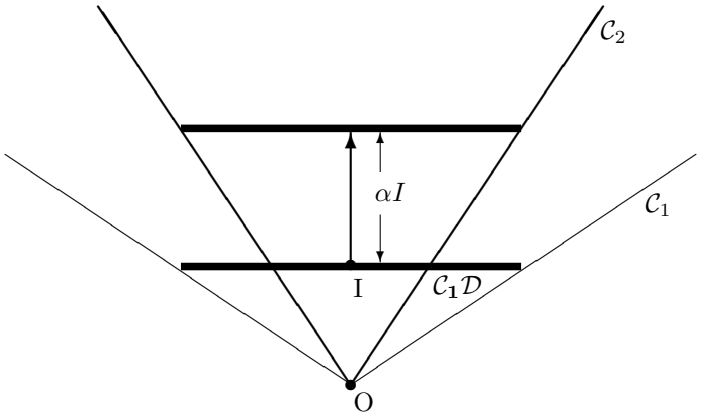


FIG. 2. *Definition of* $\alpha(\mathcal{C}_1, \mathcal{C}_2)$.

(i) $-\min_{A\in\mathcal{C}_1\mathcal{D}} \min_{X\in\mathcal{C}_2^*\mathcal{N}} \langle A, X\rangle$.
(ii) $\frac{p}{1-p}$, *where* $p = \max_{Y\in\partial\mathcal{C}_2\mathcal{D}} \min_{X\in\mathcal{C}_1^*\mathcal{N}} \langle Y, X\rangle$.
(iii) $-\min_{H\in\mathcal{H}_P} -\delta^*(H| - \mathcal{C}_2^*\mathcal{N}) + \delta(H|\mathcal{C}_1\mathcal{D})$.
(iv) $\alpha(\mathcal{C}_2^*, \mathcal{C}_1^*)$.

Proof.

(i) If we denote the number in (i) by α, we obtain that

$$\min_{A\in\mathcal{C}_1\mathcal{D}} \min_{X\in\mathcal{C}_2^*\mathcal{N}} \langle A + \alpha I, X\rangle = 0.$$

Consequently, for every $A \in \mathcal{C}_1\mathcal{D}$ and $X \in \mathcal{C}_2^*$ we have that $\langle A + \alpha I, X\rangle \geq 0$. Thus $A + \alpha I \in \mathcal{C}_2$ for every $A \in \mathcal{C}_1\mathcal{D}$. This yields $\alpha(\mathcal{C}_1, \mathcal{C}_2) \leq \alpha$.
Conversely, since $A + \alpha(\mathcal{C}_1, \mathcal{C}_2)I \in \mathcal{C}_2$ for every $A \in \mathcal{C}_1\mathcal{D}$, we get that

$$\min_{A\in\mathcal{C}_1\mathcal{D}} \min_{X\in\mathcal{C}_2^*\mathcal{N}} \langle A + \alpha(\mathcal{C}_1, \mathcal{C}_2)I, X\rangle \geq 0.$$

Consequently, $\alpha \leq \alpha(\mathcal{C}_1, \mathcal{C}_2)$.

(ii) First note that $p \leq 1$, since one may always choose $X = \frac{1}{n}I$. Since $P \neq \{\,(i,i) : i = 1, ..., n\,\}$, we have that $\partial\mathcal{C}_2\mathcal{D}$ contains only elements Y that have a nonzero off-diagonal entry (because I belongs to the interior of $\mathcal{C}_2$). But then for every such Y, an $X \in \mathcal{C}_1\mathcal{N}$ can be found so that $\langle X, Y\rangle < 1$. Since $\partial\mathcal{C}_2\mathcal{D}$ is compact, we thus obtain that $p < 1$.

Suppose that the minimum in (i) is attained at $A_{\text{opt}} \in \mathcal{C}_1\mathcal{D}$. Thus

$$\min_{X\in\mathcal{C}_2^*\mathcal{N}} \langle A_{\text{opt}} + \alpha(\mathcal{C}_1, \mathcal{C}_2)I, X\rangle = 0.$$

So $A_{\text{opt}} + \alpha(\mathcal{C}_1, \mathcal{C}_2)I \in \mathcal{C}_2$, and for every $\epsilon > 0$ there is an $X \in \mathcal{C}_2^*\mathcal{N}$ such that

$$\langle A_{\text{opt}} + \alpha(\mathcal{C}_1, \mathcal{C}_2)I - \epsilon I, X\rangle \ < 0.$$

Thus $A_{\text{opt}} + \alpha(\mathcal{C}_1, \mathcal{C}_2)I \in \partial\mathcal{C}_2$, and thus (use that $\alpha(\mathcal{C}_1, \mathcal{C}_2) \geq 0$),

$$Y := \frac{1}{1 + \alpha(\mathcal{C}_1, \mathcal{C}_2)}(A_{\text{opt}} + \alpha(\mathcal{C}_1, \mathcal{C}_2)I) \in \partial\mathcal{C}_2\mathcal{D}.$$

Consequently, since $A_{\text{opt}} \in \mathcal{C}_1$,

$$p \geq \min_{X \in \mathcal{C}_1^* \mathcal{N}} \langle Y, X \rangle \geq \frac{\alpha(\mathcal{C}_1, \mathcal{C}_2)}{1 + \alpha(\mathcal{C}_1, \mathcal{C}_2)}.$$

Conversely, suppose that the maximum in (ii) is attained at $Y_{\text{opt}} \in \partial \mathcal{C}_2 \mathcal{D}$. Then $Y_{\text{opt}} - pI \in \mathcal{C}_1$. So

$$\frac{1}{1-p}(Y_{\text{opt}} - pI) \in \mathcal{C}_1 \mathcal{D}.$$

Let $X \in \mathcal{C}_2^* \mathcal{N}$ so that $\langle Y_{\text{opt}}, X \rangle = 0$, which is possible since $Y_{\text{opt}} \in \partial \mathcal{C}_2$. Then

$$-\alpha(\mathcal{C}_1, \mathcal{C}_2) \leq \left\langle \frac{1}{1-p}(Y_{\text{opt}} - pI), X \right\rangle = \frac{-p}{1-p}.$$

(iii) This follows immediately from (i) and the observation that (iii) is equal to

$$-\min_{H \in \mathcal{C}_1 \mathcal{D}} -\delta^*(H | -\mathcal{C}_2^* \mathcal{N}).$$

(iv) Use (i) and the observation that, for positive semidefinite diagonal matrices D with $\sum D_{ii}^2 = 1$, we have that $\langle A, DXD \rangle = \langle DAD, X \rangle$. Furthermore, if $A_{ii} = 1$ then $\operatorname{trace}(DAD) = 1$. Conversely, if $X \geq 0$ and $\operatorname{trace} X = 1$ then X can be rewritten in the form DAD, with A and D as above. These observations allows one to simultaneously replace in (i) the set $\mathcal{C}_1 \mathcal{D}$ by $\mathcal{C}_1 \mathcal{N}$ and $\mathcal{C}_2^* \mathcal{N}$ by $\mathcal{C}_2^* \mathcal{D}$ without affecting its value. □

Remark. Proposition 2.1 is valid in more general cases. Analyzing the proof, one may check that, for instance, (i) and (iii) are valid when $\mathcal{C}_1 \mathcal{D}$ and $\mathcal{C}_2^* \mathcal{N}$ are compact and $\mathcal{C}_2^* \setminus \{0\} \subseteq \{ A : \operatorname{trace} A > 0 \}$. Part (iv) is valid for any pair of cones that have, in addition, the property that one can carry out the diagonal scaling conversion as performed in the proof of (iv). To avoid a cumbersome statement, we chose to present the statement as above. It is easy to check, when necessary, whether the proof applies to other cones of the reader's interest.

The numbers $\alpha(\mathcal{C}_1, \mathcal{C}_2)$ may be further interpreted in terms of eigenvalues by using the following observations. When $M = M^*$ we let $\lambda_{\min}(M)$ denote the smallest eigenvalue of M. For $M \in \mathcal{H}_P$ it is not hard to prove that

$$\begin{aligned} \min_{A \in \mathcal{A}_P^* \mathcal{N}} \langle A, M \rangle &= \min_{L \times L \subseteq P} \lambda_{\min}(M | L \times L) \\ &= \text{“smallest minimal eigenvalue among all principal submatrices of } M \text{ that lie in the pattern } P\text{.”} \end{aligned} \tag{6}$$

When the number in (6) is nonnegative we say that M is a *partial positive semidefinite matrix* with respect to P. Furthermore,

$$\begin{aligned} \min_{A \in \mathcal{X}_P \mathcal{N}} \langle A, M \rangle &= \max_{W \in \mathcal{H}_n \ominus \mathcal{H}_P} \lambda_{\min}(M + W) \\ &= \text{“ the largest possible minimal eigenvalue of a completion of } M \text{ with respect to the pattern } P\text{.”} \end{aligned} \tag{7}$$

This consequence of the Hahn-Banach separation theorem is a useful fact in some optimization problems and appears implicitly in the optimality conditions in [33] (see

also [8, Section 1.6]). When the number in (7) is nonnegative we say that M *has a positive semidefinite completion* with respect to P. Lastly,

$$\min_{A\in\mathcal{X}_P^*\mathcal{N}}\langle A, M\rangle = \lambda_{\min}(M), \tag{8}$$

which when nonnegative means that M is positive semidefinite. The inequality $\geq$ in (8) follows immediately from $M \geq \lambda_{\min}(M)I$, while the inequality $\leq$ may be obtained by choosing $A = P_{\mathcal{H}_P}(vv^*)$, where v is a normalized eigenvector of M at $\lambda_{\min}(M)$.

As an illustration of how one may apply Proposition 2.1 we state the following corollary.

COROLLARY 2.2. *Let P be a pattern and let $\nu(\cdot)$ be as defined in the introduction. Then*

$$\min_{A\in\mathcal{A}_P\mathcal{D}}\nu(A) = \tag{9}$$
$$-\max_{X\in\mathcal{X}_P\mathcal{D}}\min\left\{\lambda : \exists y_i\in\mathbb{F}^n \text{ such that } \lambda I + X = \sum y_i y_i^* \text{ and } \operatorname{supp} y_i y_i^* \subseteq P\right\},$$

where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.

Proof. Note that the left-hand side of (9) equals $-\alpha(\mathcal{A}_P, \mathcal{X}_P^*)$ and the right-hand side equals $-\alpha(\mathcal{X}_P, \mathcal{A}_P^*)$. Now apply Proposition 2.1(iv). □

Note that the right-hand side of (9) corresponds to problem 3 in the introduction.

We will return to the computation of $\alpha(\mathcal{C}_1, \mathcal{C}_2)$ for specific patterns in the following sections. In particular, we are interested in determining which patterns P give extreme cases.

Let us end this section with the observation that the number $\alpha(\mathcal{X}_P^*, \mathcal{X}_P)$ is connected to the smallest eigenvalue of the spectra of the graph P. Recall that the adjacency matrix of an undirected graph is a matrix which has a one in the entry (i, j) if vertex i is connected to vertex j, and the entry is zero otherwise. For a graph P we denote the adjacency matrix by A_P. With this notation we get the following result.

THEOREM 2.3.

$$\alpha(\mathcal{X}_P^*, \mathcal{X}_P) = -\lambda_{\min}(I + A_P), \tag{10}$$

where A_P is the adjacency matrix of the graph P.

Proof. Let $\alpha = -\lambda_{\min}(I + A_P)$. This is a lower bound for $\alpha(\mathcal{X}_P^*, \mathcal{X}_P)$, since $I + A_P \in \mathcal{X}_P^*\mathcal{D}$, and for $\epsilon > 0$ we have $(\alpha - \epsilon)I + I + A_P \notin \mathcal{X}_P$. For the converse, observe that by the Schur product theorem for positive semidefinite matrices the Schur product between any matrix from $\mathcal{X}_P^*$ with a matrix from $\mathcal{X}_P$ is positive semidefinite (see, e.g., [34, Theorem 2.1]). In particular, this means that for any $C \in \mathcal{X}_P^*\mathcal{D}$ we get

$$C \circ (I + A_P + \alpha I) = C + \alpha I \in \mathcal{X}_P.$$

Thus, $\alpha(\mathcal{X}_P^*, \mathcal{X}_P) \leq \alpha$. □

If P is a chordal graph we have the following corollary, taking into account that $\mathcal{A}_P^* = \mathcal{X}_P$ and $\mathcal{A}_P = \mathcal{X}_P^*$ in this case.

COROLLARY 2.4. *Let P be a chordal pattern. Then*

$$\alpha(\mathcal{A}_P, \mathcal{A}_P^*) = \alpha(\mathcal{A}_P, \mathcal{X}_P) = \alpha(\mathcal{X}_P^*, \mathcal{A}_P^*) = -\lambda_{\min}(I + A_P),$$

where A_P is the adjacency matrix of the graph P.

3. Extreme patterns. In this section we will consider the following problem. Given a dimension n and a combination $\mathcal{M},\mathcal{N}$ of cone types with $\mathcal{M} \in \{\mathcal{A}, \mathcal{X}^*\}$ and $\mathcal{N} \in \{\mathcal{X}, \mathcal{A}^*\}$, which pattern(s) $P \subseteq \{1,\dots,n\} \times \{1,\dots,n\}$ yields the highest values for $\alpha(\mathcal{M}_P, \mathcal{N}_P)$?

In what follows, the complete bipartite graph will appear several times. The *complete bipartite graph* $K_{m,n}$ is the graph whose vertex set consists of $m+n$ vertices. The vertices can be divided in two sets, V_1, V_2, consisting of m and n vertices, respectively, and for each vertex in V_1 there is an edge to each vertex in V_2, and $K_{m,n}$ has no other edges. The complete bipartite graph $K_{n,n}$ is extreme in the sense that it contains the maximal number of edges among the graphs with $2n$ vertices without any triangles (i.e. all their maximal cliques are of size 2); this is a part of Turan's theorem (see, e.g., [7, Chapter VI]).

We first consider the cone inclusion number $\alpha(\mathcal{X}_P^*, \mathcal{X}_P)$. In this case we are in the fortunate situation where Theorem 2.3 gives $\alpha(\mathcal{X}_P^*, \mathcal{X}_P)$ for every pattern P. This reduces the proof of the following theorem to computing the smallest minimal eigenvalue of all $n \times n$ adjacency matrices.

THEOREM 3.1. *For any pattern* $P \subseteq \{1,\dots,n\} \times \{1,\dots,n\}$,

$$\alpha(\mathcal{X}_P^*, \mathcal{X}_P) \leq \sqrt{\lfloor \frac{n}{2} \rfloor \cdot \lceil \frac{n}{2} \rceil} - 1. \tag{11}$$

Equality holds if and only if $P = K_{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$.

Proof. The inequality follows from Theorem 2.3 and Proposition 2 of [10], which gives the lower bound of the minimal eigenvalue of adjacency matrices of simple graphs on n vertices. ∎

The following proposition gives an upper bound of $\alpha(\mathcal{A}_P, \mathcal{A}_P^*)$, and also provides a proof of Theorem 1.2.

PROPOSITION 3.2. *Let* P *be a pattern in* $\{1,...,n\} \times \{1,...,n\}$. *Then*

$$-\min_{A \in \mathcal{A}_P\mathcal{D}} \min_{B \in \mathcal{A}_P\mathcal{N}} \langle A, B \rangle = \alpha(\mathcal{A}_P, \mathcal{A}_P^*) \leq n-2. \tag{12}$$

Proof. Let $A = (a_{ij}) \in \mathcal{A}_P\mathcal{D}$. Then for $i \neq j$ we have that $|a_{ij}| \leq 1$. Let E_{ij} denote the matrix with a 1 in position (i,j) and 0 elsewhere. Then

$$A + (n-2)I_n = \sum_{i<j} (|a_{ij}|(E_{ii}+E_{jj}) + a_{ij}E_{ij} + a_{ji}E_{ji}) + \sum_{i=1}^{n} \left(n - 1 - \sum_{j \neq i} |a_{ij}| \right) E_{ii}.$$

Thus we have written $A + (n-2)I_n$ as the sum of positive semidefinite rank one matrices with support in P. Consequently, $A + (n-2)I_n \in \mathcal{A}_P^*$, so by definition $\alpha(\mathcal{A}_P, \mathcal{A}_P^*) \leq n-2$. ∎

Note that we in fact showed that

$$\alpha(\mathcal{A}_P, \mathcal{D}\mathcal{D}_P) \leq n-2,$$

where $\mathcal{D}\mathcal{D}_P$ is the cone of all diagonally dominant matrices in $\mathcal{H}_P$. Clearly, $\mathcal{D}\mathcal{D}_P \subseteq \mathcal{A}_P^*$. Moreover, for any $A \in \mathcal{H}_P$,

$$\min\{\, \lambda \;:\; A + \lambda I \subseteq \mathcal{D}\mathcal{D}_P \,\} = \max_i \left(-|A_{ii}| + \sum_{j \neq i} |A_{ij}| \right).$$

Proof of Theorem 1.2. Let A be as in the statement of Theorem 1.2. Then $X := (n-1)A-(n-2)I \in \mathcal{A}_P\mathcal{D}$. Thus by Proposition 3.2 we have that $(n-1)A \in \mathcal{A}_P^*$, yielding that $(n-1)A = \sum_{i=1}^m Y_i$ for some positive semidefinite rank one Y_i's with sparsity pattern P. Thus $A = \frac{1}{n-1}\sum_{i=1}^m Y_i$, which is a sum of positive semidefinite rank one's with sparsity pattern P. ∎

The current knowledge of the extreme patterns for $\alpha(\mathcal{M}_P, \mathcal{A}_P^*)$, with $\mathcal{M} \in \{\mathcal{A}, \mathcal{X}^*\}$, can then be summarized in the following theorem. The $k \times m$ matrix with one in every entry is denoted by $J_{k\times m}$.

THEOREM 3.3. *Let $\mathcal{M}$ be any of the two cone types $\mathcal{A}$ and $\mathcal{X}^*$. For any pattern $P \subseteq \{1,\dots,n\} \times \{1,\dots,n\}$ $(n \geq 2)$,*

$$\alpha(\mathcal{M}_P, \mathcal{A}_P^*) \leq n-2. \tag{13}$$

When $P = K_{k,m}$ with $k+m=n$, we have

$$\alpha(\mathcal{M}_P, \mathcal{A}_P^*) = \sqrt{km} - 1 \leq \sqrt{\lceil \frac{n}{2} \rceil \cdot \lfloor \frac{n}{2} \rfloor} - 1. \tag{14}$$

Proof. The first statement of the theorem follows from Proposition 3.2.

To simplify notation let $\beta = \sqrt{km}$. Let

$$\begin{pmatrix} I_k & F^* \\ F & I_m \end{pmatrix} \in \mathcal{A}_P.$$

Then, by an argument similar to the proof of Proposition 3.2, one can see that

$$\begin{pmatrix} I_k & F^* \\ F & I_m \end{pmatrix} + (\beta - 1)\begin{pmatrix} I_k & 0 \\ 0 & I_m \end{pmatrix}$$

can be written as a sum of rank one positive semidefinite matrices with the specified pattern. This gives the inequality

$$\alpha(\mathcal{X}_P^*, \mathcal{A}_P^*) \leq \alpha(\mathcal{A}_P, \mathcal{A}_P^*) \leq \sqrt{km} - 1.$$

The fact that these are equalities follows by putting $F = J_{km}$. ∎

Note that from Theorem 3.3 we may conclude that the highest values for $\alpha(\mathcal{A}_P, \mathcal{A}_P^*)$ and $\alpha(\mathcal{X}_P^*, \mathcal{A}_P^*) = \alpha(\mathcal{A}_P, \mathcal{X}_P)$, with $P \subseteq \{1,...,n\} \times \{1,...,n\}$, are $\mathcal{O}(n)$. The exact values lie between $\sqrt{\lceil \frac{n}{2} \rceil \cdot \lfloor \frac{n}{2} \rfloor} - 1$ and $n-2$.

For triangle free graphs it can be seen that the number $\sqrt{\lceil \frac{n}{2} \rceil \cdot \lfloor \frac{n}{2} \rfloor} - 1$ is the extreme value of $\alpha(\mathcal{A}_P, \mathcal{X}_P) = \alpha(\mathcal{X}_P^*, \mathcal{A}_P^*)$ as follows. Recall that a graph is called *triangle free* if it does not contain any cliques of size 3 (or higher).

THEOREM 3.4. *If P is a triangle free graph, then $\alpha(\mathcal{A}_P, \mathcal{X}_P) = \lambda_{\max}(A_P) - 1$.*

Proof. Let $B \in \mathcal{A}_P\mathcal{D}$. Then, obviously, $|B| \leq A_P + I$ element wise. By a theorem of Ky Fan (see, e.g., [23, Theorem 8.2.12]), and the fact that A_P and B are Hermitian, it follows that every eigenvalue λ of B satisfies the inequality $-\lambda_{\max}(A_P) \leq \lambda - 1 \leq \lambda_{\max}(A_P)$. This implies that $B + (\lambda_{\max}(A_P) - 1)I \in \mathcal{X}_P$, which yields $\alpha(\mathcal{A}_P, \mathcal{X}_P) \leq \lambda_{\max}(A_P) - 1$.

Since P is triangle free, $I - A_P \in \mathcal{A}_P\mathcal{D}$ and $\lambda_{\min}(I - A_P) = 1 - \lambda_{\max}(A_P)$. ∎

COROLLARY 3.5. *If P is triangle free, then*

$$\alpha(\mathcal{A}_P, \mathcal{X}_P) = \alpha(\mathcal{X}_P^*, \mathcal{A}_P^*) \leq \sqrt{\lceil \frac{n}{2} \rceil \cdot \lfloor \frac{n}{2} \rfloor} - 1.$$

Equality is attained for $P = K_{\lfloor \frac{n}{2} \rfloor, \lceil \frac{n}{2} \rceil}$.

Proof. This follows by combining the proof of [11, Theorem 7.25] with Theorem 3.4. ∎

In the next section we turn our attention to the cone inclusion numbers $\alpha(\mathcal{A}_P, \mathcal{X}_P^*) = \alpha(\mathcal{X}_P, \mathcal{A}_P^*)$.

4. Applications to matrix completion problems. In this section we will present some computations of cone inclusion numbers $\alpha(\mathcal{A}_P, \mathcal{X}_P^*)$ for different types of patterns P. These results have immediate interpretations in terms of positive semidefinite matrix completion problems.

To simplify the notation in this section set

$$q(P) = \alpha(\mathcal{A}_P, \mathcal{X}_P^*) = \alpha(\mathcal{X}_P, \mathcal{A}_P^*), \tag{15}$$

$$p(P) = \frac{q(P)}{1 + q(P)}. \tag{16}$$

By Proposition 2.1 we know that

$$q(P) = -\min_{A \in \mathcal{A}_P \mathcal{D}} \min_{X \in \mathcal{X}_P \mathcal{N}} \langle A, X \rangle, \tag{17}$$

which by (7) may be interpreted in terms of a matrix completion problem for partial positive semidefinite matrices with pattern P. The interpretation is that any partial positive semidefinite matrix with pattern P, and 1's on the diagonal, has a completion whose minimal eigenvalue is at least $-q(P)$. Moreover, there exists a partial positive semidefinite partial matrix A with pattern P such that the largest possible minimal eigenvalue of any completion of A is $-q(P)$.

It can also be shown that if A is a partial Hermitian matrix with pattern P, 1's on the diagonal, and the minimal eigenvalue of every specified principal submatrix of A is greater or equal to $p(P)$, then A has a positive semidefinite completion. The number $p(P)$ corresponds to the number p introduced in Proposition 2.1(ii).

The following theorem gives the value of $q(P)$ for a new class of patterns in the positive semidefinite completion problem, which contains among others the complete bipartite graphs $K_{n,n}$.

THEOREM 4.1. *Let* P *be the pattern associated with the matrix*

$$M = \begin{pmatrix} A & C^* \\ C & B \end{pmatrix}, \tag{18}$$

where $A = \operatorname{diag}(A_1, \ldots, A_m)$, $B = \operatorname{diag}(B_1, \ldots, B_n)$, *where* $A_i = J_{p_i \times p_i}$, $1 \le i \le m$, *and* $B_i = J_{k \times k}$, $1 \le i \le n$, $p = \sum_{i=1}^m p_i$, *and* $C = J_{kn \times p}$. *If* $m \le n$, *then*

$$q_{\mathbb{R}}(P) \le q_{\mathbb{C}}(P) = -\lambda_{\min} \begin{pmatrix} A & J_{p \times k} \\ J_{k \times p} & J_{k \times k} \end{pmatrix}. \tag{19}$$

Proof. Start with an arbitrary partial correlation matrix, E, with pattern P. By filling in zeros on all the entries corresponding to the entries outside the pattern in the block diagonal matrix A, a new partial matrix, $\hat{E}$, is obtained. Each of the maximal specified principal submatrices of $\hat{E}$ is obtained by filling in zeros in a partial matrix with pattern given by the matrix

$$\tilde{X}_1 = \begin{pmatrix} A & J_{p \times k} \\ J_{k \times p} & J_{k \times k} \end{pmatrix}. \tag{20}$$

Since this pattern is chordal, it follows by Corollary 2.4 that none of the specified principal submatrices of $\hat{E}$ has an eigenvalue less than $\lambda_{\min}(\tilde{X}_1)$. Since the new partial matrix has a chordal pattern, it follows that

$$q(P) = \alpha(\mathcal{A}_P, \mathcal{X}_P^*) \leq -\lambda_{\min}(\tilde{X}_1).$$

Now we will construct complex matrices $X \in \mathcal{A}_P\mathcal{D}$, $Y \in \mathcal{X}_P\mathcal{N}$ such that $\langle X, Y\rangle = \lambda_{\min}(\tilde{X}_1)$, i.e., we show that the above inequality in the complex case is in fact an equality. Let us first assume that $m = n$.

Define the unitary matrices

$$\hat{U}_l = \operatorname{diag}\left(I_{p_1}, e^{\frac{2\pi(l-1)i}{m}} I_{p_2}, e^{\frac{2\pi(l-1)2i}{m}} I_{p_3}, \ldots, e^{\frac{2\pi(l-1)(m-1)i}{m}} I_{p_m}\right),$$
$$U_l = \operatorname{diag}(\hat{U}_l, I_k), \qquad l = 1, ..., m.$$

Note that $U_1 = I$.

Let $\tilde{Y}_1$ be a rank one positive semidefinite matrix with the properties that $\langle \tilde{X}_1, \tilde{Y}_1\rangle = \lambda_{\min}(\tilde{X}_1)$ and $\operatorname{trace}(\tilde{Y}_1) = 1$. Write

$$\tilde{Y}_1 = \begin{pmatrix} \hat{Y}_{11} & \hat{Y}_{21}^* \\ \hat{Y}_{21} & \hat{Y}_{22} \end{pmatrix},$$

using the same block decomposition of $\tilde{Y}_1$ as of $\tilde{X}_1$. Define matrices

$$\tilde{X}_l = U_l^* \tilde{X}_1 U_l,$$
$$\tilde{Y}_l = U_l^* \tilde{Y}_1 U_l.$$

It clearly follows that $\langle \tilde{X}_l, \tilde{Y}_l\rangle = \lambda_{\min}(\tilde{X}_1)$. Construct matrices Y_l by embedding the matrix $\tilde{Y}_l$ in a matrix with the same size as M by letting

$$Y_1 = \begin{pmatrix} \hat{Y}_{11} & \hat{Y}_{21}^* & 0 & \ldots \\ \hat{Y}_{21} & \hat{Y}_{22} & 0 & \ldots \\ 0 & 0 & 0 & \\ \vdots & & & \ddots \end{pmatrix},$$

$$Y_2 = \begin{pmatrix} \hat{U}_2^*\hat{Y}_{11}\hat{U}_2 & 0 & \hat{U}_2^*\hat{Y}_{21}^* & 0 & \ldots \\ 0 & \ldots & 0 & 0 & \\ \hat{Y}_{21}\hat{U}_2 & 0 & \hat{Y}_{22} & 0 & \ldots \\ 0 & 0 & 0 & & \\ \vdots & & & \ddots & \end{pmatrix},$$

$$\vdots$$

$$Y_m = \begin{pmatrix} \hat{U}_m^*\hat{Y}_{11}\hat{U}_m & 0 & \ldots & 0 & \hat{U}_m^*\hat{Y}_{21}^* \\ 0 & \ldots & \ldots & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & \ldots & \ldots & \ldots & 0 \\ \hat{Y}_{21}\hat{U}_m & 0 & \ldots & 0 & \hat{Y}_{22} \end{pmatrix}.$$

Define

$$Y = \frac{1}{m}\sum_{l=1}^{m} Y_l$$

and

$$X = \begin{pmatrix} A & \hat{U}_1^* J_{p\times k} & \hat{U}_2^* J_{p\times k} & \dots & \hat{U}_m^* J_{p\times k} \\ J_{k\times p}\hat{U}_1 & J_{k\times k} & 0 & \dots & 0 \\ J_{k\times p}\hat{U}_2 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ J_{k\times p}\hat{U}_m & 0 & \dots & 0 & J_{k\times k} \end{pmatrix}.$$

The $\hat{U}_l$'s are chosen such that the entries off the block diagonal given by the matrix A cancel in the sum $\sum_{i=1}^m \hat{U}_i^* \hat{Y}_{11} \hat{U}_i$. Therefore $Y \in \mathcal{X}_P\mathcal{N}$.

Moreover, $X \in \mathcal{A}_P\mathcal{D}$ and

$$\begin{aligned} \langle X, Y\rangle &= \frac{1}{m}\sum_{l=1}^m \langle X, Y_l\rangle \\ &= \frac{1}{m}\sum_{l=1}^m \langle \tilde{X}_l, \tilde{Y}_l\rangle \\ &= \lambda_{\min}(\tilde{X}_1). \end{aligned}$$

This completes the proof if $m = n$. If $n \geq m$, one can extend X and Y by adding extra zeroes. ∎

COROLLARY 4.2. *For the complete bipartite graphs we have* $q_{\mathbb{C}}(K_{n,n}) = \sqrt{n} - 1$, *and if* $m < n$ *then* $q_{\mathbb{C}}(K_{m,n}) = \sqrt{m} - 1$. *If* n *is a power of* 2, $q_{\mathbb{R}}(K_{n,n}) = \sqrt{n} - 1$.

Proof. By Theorem 4.1 it follows that

$$q_{\mathbb{C}}(K_{m,n}) = -\lambda_{\min}\begin{pmatrix} I_m & J_{m\times 1} \\ J_{1\times m} & 1 \end{pmatrix} = \sqrt{m} - 1.$$

If $n = 2^k$, let F be the real symmetric unitary matrix obtained using the Kronecker product by setting

$$F = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}^{\otimes k}.$$

Then we may choose

$$\begin{aligned} X &= \begin{pmatrix} I & -\sqrt{n}F \\ -\sqrt{n}F & I \end{pmatrix}, \\ Y &= \frac{1}{2n}\begin{pmatrix} I & F \\ F & I \end{pmatrix} \end{aligned}$$

in the last part of the proof of Theorem 4.1. This choice yields that $q_{\mathbb{R}}(K_{n,n}) = \sqrt{n} - 1$. ∎

Proof of Theorem 1.1. This follows immediately from Corollary 4.2. ∎

Necessary and sufficient conditions for the existence of positive semidefinite completions for partial matrices with the completely bipartite graphs can be stated as follows in terms of contractivity of an inflated Schur map. When $A \in \mathbb{C}^{n\times m}$, we denote by $\phi_A : \mathbb{C}^{n\times m} \to \mathbb{C}^{n\times m}$ the Schur map $X \to A \circ X$, where $\circ$ denotes the Schur (or Hadamard) product for matrices. If we endow $\mathbb{C}^{n\times m}$ with the spectral

norm (the largest singular value), we may rephrase a result by Haagerup [20] (see also [34, Lemma 3.1] and [30, Theorem 3.1(a)]) as follows:

$$\phi_A \text{ is a contraction} \iff \begin{pmatrix} I & A \\ A^* & I \end{pmatrix} \in \mathcal{X}^*_{K_{n,m}}.$$

When we use this correspondence, Corollary 4.2 is equivalent to the following. The $k \times k$ *Fourier matrix* is given by

$$F = \left(e^{\frac{2\pi i pq}{k}}\right)_{p,q=0}^{k-1}.$$

COROLLARY 4.3. *Let* $A = (a_{ij})_{i=1,j=1}^{n,m} \in \mathbb{C}^{n\times m}$. *Then*

$$\min_{\|\phi_A\|=1} \max_{i,j} |a_{ij}| = \frac{1}{\sqrt{\min\{n,m\}}}.$$

The extreme case is obtained when A *contains a* $p \times p$ *Fourier matrix, with* $p = \min\{n,m\}$, *and the remaining rows or columns are zero.*

Proof. All specified principal submatrices of a matrix $A \in \mathcal{X}^*_{K_{n,m}}$ are of the form

$$\begin{pmatrix} 1 & \overline{a_{ij}} \\ a_{ij} & 1 \end{pmatrix}.$$

By Corollary 4.2 and the correspondence (16) it follows that the smallest eigenvalue of the principal matrices must be at least $1 - \frac{1}{\sqrt{\min\{m,n\}}}$ if we have $\|\phi_A\| = 1$. This holds if $|a_{ij}| \le \frac{1}{\sqrt{\min\{m,n\}}}$. By the proof of Corollary 4.2, we know that we get equality when A is the $p \times p$ Fourier matrix, possibly adjoined with zero rows or columns. ∎

In recent years some results have been obtained that give conditions for the existence of positive semidefinite completions for nonchordal patterns. We have taken advantage of these results to compute some cone inclusion numbers. The first of these results concerns the cycles.

If the graph G is a cycle, conditions for the existence of a positive semidefinite completion of a real partial, positive, semidefinite matrix were found by Fiedler [16]. His result was expressed in the language of completion problems by Barrett, Johnson, and Tarazaga [5], and from their result we can compute $p_{\mathbb{R}}(C_n)$, where C_n is an n-cycle.

THEOREM 4.4. *Let* C_n *be an* n*-cycle,* $n \ge 4$. *Then* $p_{\mathbb{R}}(C_n) = 1 - \cos(\pi/n)$.

Proof. We are following the notation used in [5, Corollary 1, p. 19]. A real partial correlation matrix specified on an n-cycle may be written in the form (doing a permutation if necessary)

$$C = \begin{pmatrix} 1 & \cos\theta_{i_1} & & & \cos\theta_{i_n} \\ \cos\theta_{i_1} & 1 & \cos\theta_{i_2} & ? & \\ & \cos\theta_{i_2} & 1 & \ddots & \\ & ? & \ddots & \ddots & \cos\theta_{i_{n-1}} \\ \cos\theta_{i_n} & & & \cos\theta_{i_{n-1}} & 1 \end{pmatrix}.$$

Here the θ_i's may be numbered such that $0 \le \theta_n \le \theta_{n-1} \le \cdots \le \theta_2 \le \theta_1 \le \pi$. Then the partial matrix C has a positive semidefinite completion if and only if

$$\sum_{i=1}^{k} \theta_i \le (k-1)\pi + \sum_{i=k+1}^{n} \theta_i, \tag{21}$$

for k odd. A simple computation shows that (21) is satisfied if

$$\frac{\pi}{n} \leq \theta_n \leq \cdots \leq \theta_1 \leq \left(1 - \frac{1}{n}\right)\pi. \tag{22}$$

Since all specified principal submatrices of C are 2×2 matrices, it follows easily that (22) is satisfied if $p_{\mathbb{R}}(C_n) \geq 1 - \cos(\pi/n)$. On the other hand, it is not difficult to see that for a certain choice of θ_i's, we get an example for which the smallest eigenvalue of a specified principal submatrix of C is $1 - \cos(\pi/n)$, and there is equality in (21). ∎

For certain graphs, a necessary and sufficient condition for the existence of a real partial positive definite completion is that the cycle condition given in [5] is met for all cycles of the graph. The graphs with this property are classified in [4] and are called *cycle completable graphs.* A *minimal cycle* in a graph G is a cycle with no chord.

We have the following result for the cycle completable graphs.

THEOREM 4.5. *Let G be a cycle completable graph and let C_n be its shortest minimal cycle of length 4 or more. Then $p_{\mathbb{R}}(G) = p_{\mathbb{R}}(C_n)$.*

Proof. Since C_n is an induced subgraph of G, it clearly follows that $p(G) \geq p(C_n)$.

Now let A be any partial positive semidefinite matrix with pattern G. Let its specified principal submatrices be denoted by A_i. The theorem follows by proving that $\lambda_{\min}(A_i) \geq p(C_n)$ for every A_i ensures the existence of a positive semidefinite completion.

Now

$$\lambda_{\min}\left(\begin{pmatrix} 1 & a_{ij} \\ a_{ji} & 1 \end{pmatrix}\right) \geq p(C_n)$$

implies $|a_{ij}| \leq 1 - p(C_n)$. This holds for every specified entry of A. Since C_n is the shortest cycle in G, it follows by Theorem 4.4 that the cycle condition is met for any cycle in G, and since A is also partial positive semidefinite it clearly has a positive semidefinite completion. ∎

Although the paper [4] considers only the real positive definite completion problem, its characterization of the cycle completable graphs makes it possible to obtain an upper bound for p in the complex case.

THEOREM 4.6. *Let G be a cycle completable graph. Then $p_{\mathbb{C}}(G) \leq 1 - \frac{\sqrt{2}}{2}$.*

Proof. If G is a cycle completable graph, it has a 3-clique chordal supergraph, [4, Theorem 3]. This means that it is possible to add edges to G to obtain a chordal graph G' in such a way that all maximal cliques in G' not present in G are of size 3 (or lower).

A partial positive semidefinite matrix M with pattern G has a positive semidefinite completion if it is possible to fill in the entries corresponding to edges in $G' \setminus G$ such that the extended partial matrix is partial positive semidefinite. The new maximal cliques are 3×3 matrices, and if the smallest eigenvalues of a specified principal submatrix of M is greater or equal to $1 - \frac{\sqrt{2}}{2}$, then $|m_{ij}| \leq \frac{\sqrt{2}}{2}$, and the new 3×3 principal submatrices can be constructed by filling in zero in the unspecified entries. An easy calculation shows that

$$\begin{pmatrix} 1 & m_{ij} & 0 \\ \overline{m_{ij}} & 1 & m_{jk} \\ 0 & \overline{m_{jk}} & 1 \end{pmatrix} \geq 0. \qquad \blacksquare$$

An interesting question is to find operations on the graph G that immediately give information about $p(G)$. One such result follows immediately from Theorem 4.4, in [22]. This result considers the case where G is a graph with a subset $S \subset V(G)$, with the property that G restricted to S is a clique, and S is also a cut set of G. (A cut set of G is a subset of the vertices of G with the property that removing these vertices from the graph disconnects G.) In this case

$$q(G) = \max\{\, q(G(S_1 \cup S)), \ldots, q(G(S_r \cup S)) \,\},$$

where $G(S_i)$ denotes each of the connected components (with vertex set S_i) of G after removing the vertices S, and $G(S_i \cup S)$ denotes G restricted to the vertices $S_i \cup S$.

Another result concerning $q(G)$ which can be expressed in terms of a graph theoretic property of G is the following.

THEOREM 4.7. *Let G be a graph. Suppose v is a vertex of G connected to all the other vertices in G. Then $p(G \setminus v) = p(G)$.*

To prove Theorem 4.7 we need the following auxiliary result.

LEMMA 4.8. *Let B be an $n \times n$ correlation matrix, $0 \leq t \leq 1$, and x an $n \times 1$ vector. Then*

$$\lambda_{\min}\left(\begin{pmatrix} 1 & x^* \\ x & B \end{pmatrix}\right) \geq 1 - t$$

implies that

$$\lambda_{\min}(D^{(-1/2)}(B - xx^*)D^{(-1/2)}) \geq 1 - t,$$

where $D = \operatorname{diag}(B - xx^*) = \operatorname{diag}(I - xx^*)$, and $D^{(-1/2)}$ denotes the Moore-Penrose inverse of $D^{\frac{1}{2}}$.

Proof. If $t = 1$, the result follows by a straightforward use of the Schur complement together with the fact that $*$congruence preserves positive semidefiniteness. If $t = 0$, then it follows from the assumptions that $B = I$ and $x = 0$.

Now, assuming $0 < t < 1$, it follows that $|x_i| < 1$, and the matrix D is invertible. By a standard application of Schur complements, we have that

$$\begin{pmatrix} 1 & x^* \\ x & B \end{pmatrix} \geq (1 - t)\begin{pmatrix} 1 & 0 \\ 0 & I \end{pmatrix}$$

if and only if

$$B - \frac{xx^*}{t} + (t - 1)I \geq 0.$$

Let J be the $n \times n$ matrix which is one for every entry. Now, we add the positive semidefinite matrix $xx^* \circ ((\frac{1}{t} - 1)J + (1 - t)I)$ to the left-hand side of the inequality above, and manipulate this new inequality until we get the desired result, as follows:

$$\begin{aligned} B - \frac{xx^*}{t} + (t - 1)I + xx^* \circ \left(\left(\frac{1}{t} - 1\right)J + (1 - t)I\right) &\geq 0, \\ B - xx^* + (t - 1)I + \operatorname{diag}((1 - t)xx^*) &\geq 0, \\ B - xx^* &\geq (1 - t)\operatorname{diag}(I - xx^*), \\ D^{(-1/2)}(B - xx^*)D^{(-1/2)} &\geq (1 - t)I. \end{aligned}$$

∎

Proof of Theorem 4.7. The fact that $p(G) \geq p(G \setminus v)$ is obvious. To obtain the other inequality, first note that any completion problem on the pattern G can be reduced to a completion problem on the pattern $G \setminus v$ by taking a Schur complement eliminating the vertex v. Now make the observation that if B is a partial correlation matrix, then the matrix $D^{(-1/2)}(B - xx^*)D^{(-1/2)}$ in Lemma 4.8 will be a partial correlation matrix (with possibly some zero rows and columns adjoined). The proposition then follows, since the lemma implies that for a partial matrix with pattern G, and for which every specified principal submatrix has minimum eigenvalue larger than $1-t$, the minimum eigenvalue of any specified principal submatrix for the corresponding problem with pattern $G \setminus v$ is also bounded below by $1 - t$. ∎

Let W_n denote the n-wheel, the graph obtained from the $n-1$ cycle C_{n-1} by adjoining one new vertex which is connected to every vertex in C_{n-1}. Then as a corollary to Theorem 4.7 we have the following.

COROLLARY 4.9. *If $n \geq 5$, then $p(W_n) = p(C_{n-1})$.*

To address the question of the highest value for $\alpha(\mathcal{A}_P, \mathcal{X}_P^*)$, we next present the following conjecture. For n odd, let $\hat{K}_{\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor}$ denote the graph obtained by adding one edge in the largest vertex set in $K_{\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor}$.

CONJECTURE 4.10. *For any pattern $P \subseteq \{1, \ldots, n\} \times \{1, \ldots, n\}$ (with $n \geq 4$),*

$$\alpha(\mathcal{A}_P, \mathcal{X}_P^*) \leq \alpha(\mathcal{A}_{Q_n}, \mathcal{X}_{Q_n}^*), \tag{23}$$

where

$$Q_n = \begin{cases} K_{\frac{n}{2}, \frac{n}{2}} & \text{when } n \text{ is even,} \\ \hat{K}_{\lceil \frac{n}{2} \rceil, \lfloor \frac{n}{2} \rfloor} & \text{when } n \text{ is odd.} \end{cases} \tag{24}$$

Note that from Theorem 4.1 it follows that the cone inclusion number $\alpha(\mathcal{A}_{Q_n}, \mathcal{X}_{Q_n}^*)$ is given by

$$\alpha_{\mathbb{C}}(\mathcal{A}_{Q_n}, \mathcal{X}_{Q_n}^*) = \begin{cases} \sqrt{\frac{n}{2}} - 1 & \text{when n is even,} \\ -\lambda_{\min} \begin{pmatrix} J_{2\times 2} & 0 & J_{\lceil \frac{n}{2} \rceil \times 1} \\ 0 & I_{\lceil \frac{n}{2} \rceil - 2} & \\ & J_{1 \times \lceil \frac{n}{2} \rceil} & 1 \end{pmatrix} & \text{when n is odd.} \end{cases} \tag{25}$$

When Q_n is the complete bipartite graph $K_{\frac{n}{2}, \frac{n}{2}}$, the cone inclusion number given in (25) is obtained, for instance, when

$$A = \begin{pmatrix} I_{\frac{n}{2}} & F \\ F^* & I_{\frac{n}{2}} \end{pmatrix} \in \mathcal{A}_{Q_n} \mathcal{D},$$

$$X = \frac{1}{n} \begin{pmatrix} I_{\frac{n}{2}} & -\frac{1}{\sqrt{\frac{n}{2}}} F \\ -\frac{1}{\sqrt{\frac{n}{2}}} F^* & I_{\frac{n}{2}} \end{pmatrix} \in \mathcal{X}_{Q_n}^* \mathcal{N},$$

where F is the $\frac{n}{2} \times \frac{n}{2}$ Fourier matrix.

Let us show that

$$\alpha(\mathcal{A}_{K_{k,k}}, \mathcal{X}_{K_{k,k}}^*) < \alpha(\mathcal{A}_{\hat{K}_{k+1,k}}, \mathcal{X}_{\hat{K}_{k+1,k}}^*) < \alpha(\mathcal{A}_{K_{k+1,k+1}}, \mathcal{X}_{K_{k+1,k+1}}^*).$$

The left- and right-hand sides of the inequality are computed in Corollary 4.2. The inequalities follow by computing an estimate of the middle expression.

To compute this estimate, we use the fact that the characteristic polynomial of

$$\begin{pmatrix} J_{2\times 2} & 0 & J_{(k+1)\times 1} \\ 0 & I_{k-1} & \\ J_{1\times(k+1)} & & 1 \end{pmatrix} - I_{k+2}$$

is $(\lambda^3 - \lambda^2 - (k+1)\lambda + (k-1))(\lambda+1)\lambda^{k-2}$. (To show this, use induction and [11, Theorem 2.11].)

Let $P_{k+1}(\lambda) = \lambda^3 - \lambda^2 - (k+1)\lambda + (k-1)$, which is the third degree factor of the above characteristic polynomial. Since $P_{k+1}(-\sqrt{k+1}) = -2$, $P_{k+1}(-\sqrt{k}) = \sqrt{k} - 1$, $P_{k+1}(1) = -2$, and $\lim_{x\to\infty} P_{k+1}(x) = \infty$, it follows that

$$\sqrt{k} - 1 < \alpha\left(\mathcal{A}_{\hat{K}_{k+1,k}}, \mathcal{X}^*_{\hat{K}_{k+1,k}}\right) < \sqrt{k+1} - 1.$$

We end this section by proving that Conjecture 4.10 is true for $n \leq 5$.

As there is only one nonchordal graph with four vertices, namely $K_{2,2}$, our conjecture is trivially true in that case.

The list of nonchordal graphs of size 5 given in [4] includes six different graphs. These include $K_{2,3}$, C_5, and two other graphs which are also shown to be cycle completable. Therefore, by Theorem 4.6 it holds that $\alpha(\mathcal{A}_P, \mathcal{X}^*_P) \leq \sqrt{2} - 1$ for these graphs.

One of the two remaining graphs is the 5-wheel, W_5. By Corollary 4.9, it follows that $\alpha(\mathcal{A}_{W_5}, \mathcal{X}^*_{W_5}) = \sqrt{2} - 1$.

To compute the cone inclusion number for the last graph, which is $\hat{K}_{3,2}$, we use Theorem 4.1. The matrix for this graph is

$$A_{\hat{K}_{3,2}} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Now apply the theorem with $A = \operatorname{diag}(J_{2\times 2}, 1)$, $B = I_2$, and $C = J_{2\times 3}$ which gives

$$\alpha(\mathcal{A}_{\hat{K}_{3,2}}, \mathcal{X}^*_{\hat{K}_{3,2}}) = -\lambda_{\min} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \approx 0.4812.$$

By using the construction given in the proof of Theorem 4.1, we obtain

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 & 1 \end{pmatrix}$$

as a matrix which realizes this cone inclusion number.

5. The contractive case. Arveson [3] provided an elegant formula for the distance of a bounded linear operator on a Hilbert space $\mathcal{H}$ to a nest algebra in the C*-algebra of bounded linear operators on $\mathcal{H}$. For a finite nest we may restate this as follows: The minimal (operator) norm of a completion of

$$\begin{pmatrix} A_{11} & & ? \\ \vdots & \ddots & \\ A_{n1} & \cdots & A_{nn} \end{pmatrix} \tag{26}$$

is

$$\max_{i=1,\ldots,n} \left\| \begin{pmatrix} A_{i1} & \cdots & A_{ii} \\ \vdots & & \vdots \\ A_{n1} & \cdots & A_{ni} \end{pmatrix} \right\|. \tag{27}$$

The result holds also in the case where the resulting operator matrix acts between different Hilbert spaces. It is obvious that, for any pattern of known and unknown entries, the minimal norm of a completion is always bounded below by the maximum among the norms of specified submatrices. As Arveson's distance formula implies, these two numbers are the same for triangular patterns. In [25] it was shown that, in fact, direct sums of triangular patterns are the only ones for which this equality holds. To study what happens with other patterns, the following notion of a distance constant was introduced: For a partial matrix A with pattern $K \subseteq \{1,\ldots,n\} \times \{1,\ldots,m\}$, we introduce

$$\mu(A) = \min_{B \text{ completion of } A} \| B \|$$

and

$$\varrho(A) = \max_{I \times J \subseteq K} \| A \mid I \times J \| .$$

Note that in this section the pattern K does not need to be symmetric nor necessarily contain the main diagonal. We say that $c(K) = c \geq 1$ is the *distance constant* for pattern K if, for every partial matrix A with pattern K, we have that

$$\mu(A) \leq c\varrho(A),$$

and c is the smallest number with this property. From [3] we obtain that if K is triangular, then $c(K) = 1$, and from [25] we obtain that if K is not (permutation equivalent to) a direct sum of triangular patterns, then $c(K) > 1$. The recent paper [14] provides some distance constants for some low dimensional cases, and in addition, provides a guide to the literature on the subject.

The description of the pattern in terms of a graph in the contractive case is different from the Hermitian case. In the contractive case we use the following correspondence between a pattern and a bipartite graph. For an $n \times m$ partial matrix A, the set of vertices of the corresponding bipartite graph G is the union of two disjoint subsets $U = \{u_1, u_2, \ldots, u_n\}$ and $V = \{v_1, v_2, \ldots, v_m\}$. An edge (u_p, v_q) occurs in G if and only if the (p, q) entry of A is specified. G has no edges among the vertices in V nor among the vertices in U. In this section we will use both this description of a pattern and the description used for partial Hermitian matrices. It should follow from the context which one we are using.

The following type of question is of particular interest: Given a sequence of patterns $K_n \subseteq \{1,\dots,n\}\times\{1,\dots,n\}$, is $\lim_{n\to\infty} c(K_n)$ finite or not? An important example is the case when $T_{n,n} \subseteq \{1,\dots,n^2\}\times\{1,\dots,n^2\}$ corresponds to the pattern for which the minimal norm completion problem is equivalent to determining the distance to $\mathcal{T}_n \otimes \mathcal{T}_n$ in $\mathbb{C}^{n^2}$, where $\mathcal{T}_n$ is the nest algebra of upper triangular $n \times n$ matrices. The patterns for $n = 2$ and $n = 3$ are

$$T_{2,2} = \begin{pmatrix} ? & ? & ? & ? \\ & ? & & ? \\ & & ? & ? \\ & & & ? \end{pmatrix} \tag{28}$$

and

$$T_{3,3} = \begin{pmatrix} ? & ? & ? & ? & ? & ? & ? & ? & ? \\ & ? & ? & & ? & ? & & ? & ? \\ & & ? & & & ? & & & ? \\ & & & ? & ? & ? & ? & ? & ? \\ & & & & ? & ? & & ? & ? \\ & & & & & ? & & & ? \\ & & & & & & ? & ? & ? \\ & & & & & & & ? & ? \\ & & & & & & & & ? \end{pmatrix}. \tag{29}$$

Let us view these problems in a broader context and relate them to the previous sections. To do this we use H. Wielandt's [41] convenient equivalence

$$\begin{pmatrix} I & A \\ A^* & I \end{pmatrix} \geq 0 \qquad \Leftrightarrow \qquad \| A \| \leq 1. \tag{30}$$

For $K \subseteq \{1,\dots,n\}\times\{1,\dots,m\}$, denote by $\mathcal{M}_K$ the matrices with support in K, i.e.,

$$\mathcal{M}_K = \{ C \in \mathbb{F}^{n\times m} : C_{ij} = 0, (i,j) \notin K \}.$$

We say that $C \in \mathcal{M}_K$ has a *K-subordinate isometric/co-isometric factorization* if there exist isometries $U \in \mathbb{F}^{n\times k}$ and $V \in \mathbb{F}^{m\times k}$ such that

$$C = UV^*, \tag{31}$$

with the property that for all $(i,j) \notin K$ and all $p \in \{1,\dots,k\}$ we have $u_{ip} = 0$ or $v_{jp} = 0$. From $C \in \mathcal{M}_K$ and (31), we obtain that for $(i,j) \notin K$,

$$0 = c_{ij} = \sum_{p=1}^{k} u_{ip}\bar{v}_{jp}.$$

Thus the requirement that $u_{ip}\bar{v}_{jp} = 0$ for all $p = 1,\dots,k$ is a stronger one.

Denote $K^c = \{1,\dots,n\}\times\{1,\dots,m\}\backslash K$. Introduce the following closed convex sets in $\mathcal{M}_K$:

$$\mathcal{C}_K = \{ C \in \mathcal{M}_K : C \text{ has a } K\text{-subordinate isometric/co-isometric factorization} \},$$
$$\mathcal{D}_K = \{ C \in \mathcal{M}_K : ||C|| \leq 1 \},$$

$$\begin{aligned}\mathcal{E}_K &= \{\, C \in \mathcal{M}_K : \text{ there is a } W \in \mathcal{M}_{K^c} \text{ such that } ||C+W|| \le 1 \,\}\\ &= \text{“all } C \in \mathcal{M}_K \text{ that have a contractive completion,”}\\ \mathcal{F}_K &= \{\, C \in \mathcal{M}_K : ||C|_{P\times Q}|| \le 1 \text{ for all } P \times Q \subseteq K \,\}\\ &= \text{“all partial contractions in } \mathcal{M}_K\text{.”}\end{aligned}$$

To show that $\mathcal{C}_K$ is convex, observe that when $C = UV^*$ and $D = WY^*$ are K-subordinate isometric/co-isometric factorizations for C and D, respectively, then

$$(\alpha C + \beta D) = \begin{pmatrix}\alpha^{\frac{1}{2}}U & \beta^{\frac{1}{2}}W\end{pmatrix}\begin{pmatrix}\alpha^{\frac{1}{2}}V^* \\ \beta^{\frac{1}{2}}Y^*\end{pmatrix}$$

is a K-subordinate isometric/co-isometric factorization of $\alpha C+\beta D$. Furthermore, we have that

$$\mathcal{C}_K \subseteq \mathcal{D}_K \subseteq \mathcal{E}_K \subseteq \mathcal{F}_K. \tag{32}$$

For any pair $\mathcal{P}, \mathcal{Q}$ of sets from $\mathcal{C}_K, \mathcal{D}_K, \mathcal{E}_K, \mathcal{F}_K$, we define

$$\beta(\mathcal{P}, \mathcal{Q}) = \inf\{\, \mu : \mathcal{P} \subseteq \mu\mathcal{Q} \,\}.$$

To relate this number to cone inclusion numbers, let $\hat{\mathcal{Q}}$ denote the cone associated to $\mathcal{Q}$ defined by

$$\hat{\mathcal{Q}} = \left\{ \lambda \begin{pmatrix} I & Q \\ Q^* & I \end{pmatrix} : \lambda \ge 0, Q \in \mathcal{Q} \right\}.$$

Then, for all possible choices of $\mathcal{P},\ \mathcal{Q}$ from $\mathcal{C}_K, \mathcal{D}_K, \mathcal{E}_K, \mathcal{F}_K$, we have

$$\beta(\mathcal{P}, \mathcal{Q}) = \alpha(\hat{\mathcal{P}}, \hat{\mathcal{Q}}) + 1 := \min\{\, \lambda : \hat{\mathcal{P}}\mathcal{D} + \lambda I \subseteq \hat{\mathcal{Q}} \,\} + 1. \tag{33}$$

It should be noted that some of the β's correspond to familiar numbers, e.g.,

$$\begin{aligned}\beta(\mathcal{F}_K, \mathcal{E}_K) &= c(K) = \text{ “distance constant of } K\text{,”}\\ \beta(\mathcal{E}_K, \mathcal{D}_K) &= ||P_K||,\end{aligned}$$

where $P_K : \mathbb{F}^{n\times m} \to \mathcal{M}_K$ is the canonical projection.

As an aside, let us observe that

$$\hat{\mathcal{C}}_K \subseteq \mathcal{A}^*_{\hat{K}}, \quad \hat{\mathcal{D}}_K \subseteq \mathcal{X}_{\hat{K}}, \hat{\mathcal{E}}_K \subseteq \mathcal{X}^*_{\hat{K}}, \quad \hat{\mathcal{F}}_K \subseteq \mathcal{A}_{\hat{K}}, \tag{34}$$

where

$$\begin{aligned}\hat{K} = {} & \{\, 1,\dots,n \,\} \times \{\, 1,\dots,n \,\} \cup \{\, n+1,\dots,n+m \,\} \times \{\, n+1,\dots,n+m \,\}\\ & \cup \{\, (i,j) : (i, j-m) \in K \text{ or } (j, i-n) \in K \,\}.\end{aligned}$$

PROPOSITION 5.1. *For all $K \subseteq \{\, 1,\dots,n \,\} \times \{\, 1,\dots,m \,\}$, we have that*

$$\beta(\mathcal{F}_K, \mathcal{C}_K) \le \sqrt{\min\{n, m\}}. \tag{35}$$

Proof. Without loss of generality, $n \le m$. Let $C \in \mathcal{F}_K$, and consider

$$\begin{pmatrix} \sqrt{n}I_n & C \\ C^* & \sqrt{n}I_m \end{pmatrix} = \sum_{i=1}^{n} \begin{pmatrix} \sqrt{n}e_ie_i^* & c_ie_i^* \\ e_ic_i^* & \frac{1}{\sqrt{n}}I_m \end{pmatrix} \in \hat{\mathcal{C}}_K,$$

where e_i and c_i are the ith columns of I_n and C, respectively. Thus $\alpha(\hat{\mathcal{F}}_K\mathcal{D}, \hat{\mathcal{C}}_K) \leq \sqrt{n} - 1$, and thus (35) follows from (33). ∎

COROLLARY 5.2. *For all $K \subseteq \{1, \dots, n\} \times \{1, \dots, m\}$ and all possible choices $\mathcal{P}$ and $\mathcal{Q}$ from $\mathcal{C}_K, \mathcal{D}_K, \mathcal{E}_K, \mathcal{F}_K$, we have that*

$$\beta(\mathcal{P}, \mathcal{Q}) \leq \sqrt{\min\{n, m\}}.$$

Proof. All $\beta(\mathcal{P}, \mathcal{Q})$ are bounded above by $\beta(\mathcal{F}_K, \mathcal{C}_K)$ because of (32). Now use Proposition 5.1. ∎

COROLLARY 5.3. *For any pattern $K \subseteq \{1, \dots, n\} \times \{1, \dots, m\}$, we have*

$$c(K) \leq \sqrt{\min\{n, m\}}\ ,\ ||P_K|| \leq \sqrt{\min\{n, m\}}.$$

Let us remark that in [15], patterns $K_n \subseteq \{1, \dots, n\}^2$ are described in which n is of the form $n = 2(3^m)$ with the property that $c(K_n) \geq (\frac{n}{2})^{\log_3(\frac{3}{4}\sqrt{2})}$ (which is approximately $\mathcal{O}(n^{.0536})$). This shows that there are sequences of patterns K_n for which $\lim_{n\to\infty} c(K_n) = \infty$. It should be noticed that in this example the growth rate is much lower than the growth rate of the upper bound in Corollary 5.3. It remains to improve on this result and create (if existent), for instance, a sequence $K_n \subseteq \{1, \dots, n\} \times \{1, \dots, n\}$ for which $c(K_n) = \mathcal{O}(\sqrt{n})$.

Let $T_{n,m} \subseteq \{1, \dots, nm\} \times \{1, \dots, nm\}$ denote, as before, the pattern

$$T_{n,m} = \{(i, j) : i > j \text{ or } (i \bmod m) > (j \bmod m)\}.$$

THEOREM 5.4.

$$c(T_{n,m}) = \beta(\mathcal{F}_{T_{n,m}}, \mathcal{E}_{T_{n,m}}) \leq \lceil \log_2(\min\{n, m\}) \rceil + 1.$$

In particular,

$$c(T_{n,n}) \leq \lceil \log_2 n \rceil + 1.$$

Proof. Without loss of generality, $n \leq m$. As $c(T_{n,m}) \leq c(T_{\hat{n},m})$ when $n \leq \hat{n}$, we may assume that $n = 2^k$ for some $k = \mathbb{N}_0$. We shall perform induction on k. When $k = 0, c(T_{1,m}) = 1$ by Arveson's distance formula [3]. Suppose $c(T_{2^{k-1},m}) \leq k$, and let $C \in \mathcal{F}_{T_{2^k,m}}$. Decompose C as a 2×2 block matrix

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

with C_{ij} of size $(2^{k-1}m) \times (2^{k-1}m)$. Note that

$$\begin{aligned} C_{11}, C_{22} &\in \mathcal{F}_{T_{2^{k-1},m}}, \\ C_{21} &\in \mathcal{F}_P, \qquad C_{12} \in \mathcal{F}_Q, \end{aligned}$$

where $P = \{1, \dots, 2^{k-1}m\}^2$, and $Q \subseteq \{1, \dots, 2^{k-1}m\}^2$ is a pattern which after permutation is block triangular. In particular, $\|C_{21}\| \leq 1$. By Arveson's distance formula, there is a $W_{12} \in \mathcal{M}_{Q^c}$ such that $||C_{12} + W_{12}|| \leq 1$. By the induction hypothesis, there exist $W_{11}, W_{22} \in \mathcal{M}_{T^c_{2^{k-1},m}}$ such that

$$||C_{ii} + W_{ii}|| \leq k, i = 1, 2.$$

But then

$$\begin{pmatrix} W_{11} & W_{12} \\ 0 & W_{22} \end{pmatrix} \in \mathcal{M}_{T^c_{2^k,m}}$$

and

$$\left\| \begin{pmatrix} C_{11}+W_{11} & C_{12}+W_{12} \\ C_{21} & C_{22}+W_{22} \end{pmatrix} \right\| \leq k+1.$$

This shows that $c(T_{2^k,m}) \leq k+1$, proving the theorem. □

The question of whether

$$c(T_{n,n}) \to \infty$$

as $n \to \infty$ remains open.

Let us end this section with some remarks regarding the triangular truncation operator P_{T_n}, where

$$T_n = \{\, (i,j) : 1 \leq j < i \leq n \,\} \; (= T_{n,1}).$$

It is well known that $||P_{T_n}||$ grows like $\log n$ (see [27]). To see that

$$\beta(\mathcal{E}_{T_n}, \mathcal{D}_{T_n}) = ||P_{T_n}|| \leq \lceil \log_2 n \rceil + 1,$$

one may use that if

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

then

$$P_{T_{2^k}}(A) = \begin{pmatrix} P_{T_{2^{k-1}}}(A_{11}) & 0 \\ A_{21} & P_{T_{2^{k-1}}}(A_{22}) \end{pmatrix}.$$

So, by an induction argument, we get that

$$\begin{aligned} ||P_{T_{2^k}}(A)|| &\leq \left\| \begin{pmatrix} P_{T_{2^{k-1}}}(A_{11}) & 0 \\ 0 & P_{T_{2^{k-1}}}(A_{22}) \end{pmatrix} \right\| + \left\| \begin{pmatrix} 0 & 0 \\ A_{21} & 0 \end{pmatrix} \right\| \\ &\leq \max\{k||A_{11}||, k||A_{22}||\} + ||A_{21}|| \\ &\leq (k+1)||A||. \end{aligned}$$

In [12, Example 4.1] an example was given showing that

$$\liminf_{n\to\infty} \frac{||P_{T_n}||}{\log n} \geq \frac{4}{5\pi}.$$

Recently, in [2] it was shown that

$$\lim_{n\to\infty} \frac{||P_{T_n}||}{\log n} = \frac{1}{\pi}.$$

Some more recent bounds can be found in [29]. Finally, the book [18] has a comprehensive account of triangular truncation and triangular integral operators.

6. The Toeplitz case. A variation of the theory presented so far is to endow the matrices with some structure. A natural choice is to consider the Toeplitz case. Positive definite Toeplitz matrices are of importance in several applications. Completion problems for partial positive semidefinite Toeplitz matrices have attracted some attention due to their connection to problems in function theory and engineering. It follows as a corollary of The Carathéodory–Toeplitz interpolation theorem [9, 17, 39] that any partial positive semidefinite Toeplitz matrix specified in a band has a positive semidefinite completion.

Since a Toeplitz matrix is characterized by the property that all entries along a certain diagonal are constant (i.e., $t_{i,j} = t_{i-j}$), the natural way of describing the pattern P of a partial Hermitian Toeplitz matrix is in terms of a set on the natural numbers, where $j \in P$ if $t_j (= \overline{t_{-j}})$ is specified and $j > 0$. As before, we shall always assume that the main diagonal is specified.

Let $\mathcal{TH}_n$ denote subspace of $\mathcal{H}_n$ consisting of the $n \times n$ Hermitian Toeplitz matrices. For a pattern P we introduce the subspace

$$\mathcal{TH}_P = \{\, T \in \mathcal{TH}_n : T_{i-j} = 0 \text{ if } |i-j| \notin P \,\}.$$

In this section we consider only the two cones

$$\begin{aligned}
\mathcal{TA}_P &= \{\, A \in \mathcal{TH}_P : A \in \mathcal{A}_P \,\}, \\
\mathcal{TX}_P^* &= \{\, Y \in \mathcal{TH}_P : \text{ there is a } W \in \mathcal{TH}_n \ominus \mathcal{TH}_P \text{ such that } Y + W \geq 0 \,\}.
\end{aligned}$$

We have retained here the notation introduced in the previous sections but have added a prefix $\mathcal{T}$ to denote that we are working with Toeplitz matrices. This means that $\mathcal{TA}_P$ can be identified with the partial positive semidefinite Toeplitz matrices with pattern P, and $\mathcal{TX}_P^*$ can be identified with those partial positive semidefinite matrices which have a positive semidefinite Toeplitz completion.

We continue to use a suffix $\mathcal{D}$ to denote a restriction to the set of matrices with the main diagonal equal to the identity. Now we define

$$\begin{aligned}
q_{\mathcal{T}}(P) &= \min\{\, \lambda : \mathcal{TA}_P\mathcal{D} + \lambda I \subseteq \mathcal{TX}_P^* \,\}, \\
p_{\mathcal{T}}(P) &= \frac{q_{\mathcal{T}}(P)}{1 + q_{\mathcal{T}}(P)},
\end{aligned}$$

and recall that these numbers have the same interpretation, in terms of matrix completion problems, as $q(P)$ and $p(P)$ had in section 4.

It was proved in [24] that $q_{\mathcal{T}}(P) = 0$ if and only if

$$P = \{\, k, 2k, 3k, \ldots, mk \,\}, \qquad (k \geq 1).$$

The completion problem for the pattern $P = \{m, n\}$ is studied in [31], and using ideas from that paper it is possible to compute $p_{\mathcal{T}}^{\mathbb{C}}(\{m, n\})$. We may assume that $\gcd(m, n) = 1$.

A partial positive semidefinite Toeplitz matrix with pattern $P = \{m, n\}$ has a positive semidefinite Toeplitz completion if and only if (a_m, b_m, a_n, b_n) is located in the convex hull generated by the curve

$$w_{mn}(t) = \{\, (\cos 2\pi mt, \sin 2\pi mt, \cos 2\pi nt, \sin 2\pi nt), \quad 0 \leq t < 1 \,\}. \tag{36}$$

A description of the facial structure of the convex hull generated by the curve (36) was obtained in [38].

THEOREM 6.1. *Let* $\gcd(m,n) = 1$, $m < n$ *and* $n > 2$. *Then* $p_{\mathcal{T}}^{\mathbb{C}}(\{m,n\}) = 1 - \cos(\frac{\pi}{m+n})$.

Proof. Let the specified diagonals of the partial Hermitian Toeplitz matrix be t_m and t_n. Let us write $(t_m, t_n) \in \mathcal{TX}^*_{\{m,n\}}\mathcal{D}$ if the partial matrix has a positive semidefinite Toeplitz completion and 1's on the diagonal. With the pattern under discussion, all specified principal submatrices are 2×2 matrices, and therefore $p_{\mathcal{T}}(\{m,n\}) = 1-r$, where

$$r = \min_{(t_m,t_n)\in\mathcal{TX}^*_{\{m,n\}}\mathcal{D}} \max(|t_m|, |t_n|).$$

(t_m, t_n) must of course be located on the boundary of $\mathcal{TX}^*_{\{m,n\}}\mathcal{D}$. From [38, Theorem 1] we know that we may assume that

$$\begin{aligned}(t_m, t_n) = (&\lambda + (1-\lambda)(\cos(2p\pi t) + i\sin(2p\pi t)),\\ &\lambda + (1-\lambda)(\cos(2n\pi t) + i\sin(2n\pi t))).\end{aligned}$$

Then it follows from the fact that

$$(\lambda + (1-\lambda)p)^2 + (1-p^2)(1-\lambda)^2$$

is minimized by choosing $\lambda = 1/2$, that it is enough to find the ϕ which minimizes $\max(|t_m|, |t_n|)$, where

$$\begin{aligned}|t_m(\phi)| &= \left|\left(\frac{1}{2} + (1-\frac{1}{2})\right)(\cos(2m\pi\phi) + i\sin(2m\pi\phi))\right|\\ &= |\cos(m\pi\phi)|,\\ |t_n(\phi)| &= \left|\left(\frac{1}{2} + (1-\frac{1}{2})\right)(\cos(2n\pi\phi) + i\sin(2n\pi\phi))\right|\\ &= |\cos(n\pi\phi)|.\end{aligned}$$

The point is on the boundary of $\mathcal{TX}^*_{\{m,n\}}\mathcal{D}$ only if

$$\phi \in \left(\frac{k_1}{m}, \frac{l_1}{n}\right) \cup \left(\frac{n-l_1}{n}, \frac{m-k_1}{p}\right),$$

where l_1 and k_1 is the unique pair of integers k_1, l_1, $0 \le k_1 < m$, $1 \le l_1 < n$, such that $l_1 m - k_1 n = 1$. The two intervals are symmetric. Therefore we may make the substitution $\phi = \frac{k_1}{m} + \frac{\theta}{mn}$, where $0 \le \theta \le 1$. This gives

$$\begin{aligned}|t_m(\theta)| &= \left|\cos\left(\frac{\pi\theta}{n}\right)\right|,\\ |t_n(\theta)| &= \left|\cos\left(\frac{(\theta-1)\pi}{m}\right)\right|.\end{aligned}$$

Since $n \ge 3$, $|t_m(\frac{\pi\theta}{n})|$ is monotonically increasing, whereas $|t_n(\frac{(\theta-1)\pi}{m})|$ is monotonically decreasing when $m > 1$. The desired minimum is therefore obtained when $|t_m(\frac{\pi\theta}{n})| = |t_n(\frac{(\theta-1)\pi}{m})|$. This gives $\theta = \frac{n}{m+n}$, and $r = \cos(\frac{\pi}{m+n})$. A slight modification is needed when $m = 1$, and this is left to the reader. □

Due to the connections of the positive semidefinite Toeplitz completion problem with other fields it is possible to make different interpretations of $p_{\mathcal{T}}(P)$ in this case. Here is an example.

A real partial Hermitian Toeplitz matrix with pattern

$$P_{2n+1} = \{\, 1, 3, 5, \ldots, 2n+1 \,\}$$

has a positive semidefinite Toeplitz completion if $(t_1, t_3, t_5, \ldots, t_{2n+1})$ is located in the convex hull generated by the curve

$$(\cos\theta, \cos 3\theta, \cos 5\theta, \ldots, \cos(2n+1)\theta).$$

Let us denote this convex set as C_{2n+1}. Since in this case all specified principal submatrices are of size 2×2, the number $1 - p_T^{\mathbb{R}}(P_{2n+1})$ is equal to the minimal distance in the l_∞-norm from the origin to the boundary of C_{2n+1}. An interesting problem is to determine $\lim_{n\to\infty} 1 - p_T^{\mathbb{R}}(P_{2n+1})$.

Toeplitz matrices are also of interest in the contractive case. In [26] all patterns are described for which a partial Toeplitz contraction always allows a contractive Toeplitz completion, under the assumption that the specified entries occur in consecutive diagonals. As opposed to the non-Toeplitz case, there exist examples of triangular patterns for which not every partial Toeplitz contraction has a contractive Toeplitz completion, see, e.g., [40, Example 7.1]. A partial result in which the pattern consists of nonconsecutive diagonals is presented in [26]. The patterns which allow every partial Toeplitz contraction to have a triangular contractive Toeplitz completion were obtained in [32].

Acknowledgments. Geir Nævdal would like to express his gratitude to Professor Charles R. Johnson for helping to arrange his visits to the College of William and Mary and for extending kind hospitality. He also wishes to thank the Mathematics Department at the College for its hospitality.

Both authors gratefully acknowledge Professor C. R. Johnson for suggesting the problems which gave rise to Theorems 4.4 and 6.1 and for a useful discussion leading to the proof of Theorem 4.5.

That the paper [10] contained the result needed to prove this theorem was pointed out to us by D. M. Cvetković and S. Simic.

Finally, the authors also wish to thank the referees for their constructive criticism.

REFERENCES

[1] J. Agler, J. W. Helton, S. McCullough, and L. Rodman, *Positive semidefinite matrices with a given sparsity pattern*, Linear Algebra Appl., 107 (1988), pp. 101–149.

[2] J. R. Angelos, C. C. Cowen, and S. K. Narayan, *Triangular truncation and finding the norm of a Hadamard multiplier*, Linear Algebra Appl., 170 (1992), pp. 117–135.

[3] W. B. Arveson, *Interpolation problems in nest algebras*, J. Funct. Anal., 20 (1975), pp. 208–233.

[4] W. W. Barrett, C. R. Johnson, and R. Loewy, *The Real Positive Definite Completion Problem: Cycle Completability*, Vol. 122 of Mem. Amer. Math. Soc. 122, Amer. Math. Soc., Providence, RI, 1996.

[5] 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.

[6] A. Berman, *Cones, Matrices and Mathematical Programming*, Springer-Verlag, Berlin, New York, 1973.

[7] B. Bollobas, *Extremal Graph Theory*, Academic Press, New York, 1978.

[8] S. Boyd and L. El Ghaoui, *Method of centers for minimizing generalized eigenvalues*, Linear Algebra Appl., 188/189 (1993), pp. 63–111.

[9] C. Carathéodory, *Über den Variabilitätsbereich der Fourier'schen Konstanten von positiven harmonischen Funktionen*, Rend. Circ. Mat. Palermo, XXXII (1911), pp. 193–217.

[10] G. Constantine, *Lower bounds on the spectra of symmetric matrices with nonnegative entries*, Linear Algebra Appl., 65 (1985), pp. 171–178.
[11] D. M. Cvetkovic, M. Doob, and H. Sachs, *Spectra of Graphs*, VEB Deutsher Verlag der Wissenschaften, Berlin, 1980.
[12] K. R. Davidson, *Nest Algebras*, Pitman Research Notes in Mathematics 191, Longman Scientific and Technical, Essex, UK, 1988.
[13] K. R. Davidson, *Finite dimenison problems in operator theory*, in The Gohberg Anniversary Collection Volume I: The Calgary Conference and Matrix Theory Papers, Oper. Theory Adv. Appl. 40, H. Dym, S. Goldberg, M. Kaashoek, and P. Lancaster, eds., Birkhäuser, Basel, 1989, pp. 187–201.
[14] K. R. Davidson and M. S. Ordower, *Some exact distance constants*, Linear Algebra Appl., 208/209 (1994), pp. 37–55.
[15] K. R. Davidson and S. Power, *Failure of the distance formula*, J. London Math. Soc. (2), 32 (1984), pp. 157–165.
[16] M. Fiedler, *Matrix inequalities*, Numer. Math., 9 (1966), pp. 109–119.
[17] C. Foias and A. Frazho, *The Commutant Lifting Approach to Interpolation Problems*, Operator Theory: Advances and Applications 44, Birkhäuser Verlag, Basel, 1990.
[18] I. C. Gohberg and M. G. Krein, *Theory and Applications of Volterra Operators in Hilbert Space*, Transl. Math. Monographs 24, Amer. Math. Soc., Providence, RI, 1970.
[19] B. 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.
[20] U. Haagerup, *Decomposition of completely bounded maps on operator algebras*, manuscript.
[21] J. W. Helton, D. Lam, and H. J. Woerdeman, *Sparsity patterns with high rank extremal positive semidefinite matrices*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 299–312.
[22] J. W. Helton, S. Pierce, and L. Rodman, *The ranks of extremal positive semidefinite matrices with given sparsity pattern*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 407–423.
[23] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
[24] C. R. Johnson, M. Lundquist, and G. Nævdal, *Positive definite Toeplitz completions*, J. London Math. Soc., to appear.
[25] C. R. Johnson and L. Rodman, *Completion of partial matrices to contractions*, J. Funct. Anal., 69 (1986), pp. 260–267.
[26] C. R. Johnson and L. Rodman, *Completion of Toeplitz partial contractions*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 159–167.
[27] S. Kwapien and A. Pelczynski, *The main triangle projection in matrix spaces and its application*, Studia Math., 34 (1970), pp. 43–68.
[28] H. Lev-Ari, S. R. Parker, and T. Kailath, *Multidimensional maximum-entropy covariance extension*, IEEE Trans. Inform. Theory, 35 (1989), pp. 497–508.
[29] R. Mathias, *The Hadamard operator norm of a circulant and applications*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1152–1167.
[30] R. Mathias, *Matrix completions, norms, and Hadamard products*, Proc. Amer. Math. Soc., 117 (1993), pp. 905–918.
[31] G. Nævdal, *On a generalization of the trigonometric moment problem*, Linear Algebra Appl., 258 (1997), pp. 1–18.
[32] G. Nævdal, *On the completion of partially given triangular Toeplitz matrices to contractions*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 545–552.
[33] M. L. Overton, *On minimizing the maximum eigenvalue of a symmetric matrix*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 256–268.
[34] V. I. Paulsen, S. C. Power, and R. R. Smith, *Schur products and matrix completions*, J. Funct. Anal., 85 (1989), pp. 151–178.
[35] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[36] W. Rudin, *The extension problem for positive-definite functions*, Illinois J. Math., 7 (1963), pp. 532–539.
[37] N. Shaked-Monderer, *Extremal positive semidefinite matrices with given sparsity pattern*, Linear and Multilinear Algebra, 36 (1994), pp. 287–292.
[38] Z. Smilansky, *Convex hulls of generalized moment curves*, Israel J. Math., 52 (1985), pp. 115–128.
[39] O. Toeplitz, *Über die Fourier'sche Entwickelung positiver Funktionen*, Rend. Circ. Mat. Palermo, XXXII (1911), pp. 191–192.
[40] H. J. Woerdeman, *Strictly contractive and positive completions for block matrices*, Linear Algebra Appl., 136 (1990), pp. 63–105.
[41] H. Wielandt, *An extremum property of sums of eigenvalues*, Proc. Amer. Math. Soc., 6 (1955), pp. 106–110.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 640–663, July 1998

003

A SCHUR–FRÉCHET ALGORITHM FOR COMPUTING THE LOGARITHM AND EXPONENTIAL OF A MATRIX*

C. S. KENNEY† AND A. J. LAUB‡

Abstract. The Schur–Fréchet method of evaluating matrix functions consists of putting the matrix in upper triangular form, computing the scalar function values along the main diagonal, and then using the Fréchet derivative of the function to evaluate the upper diagonals. This approach requires a reliable method of computing the Fréchet derivative. For the logarithm this can be done by using repeated square roots and a hyperbolic tangent form of the logarithmic Fréchet derivative. Padé approximations of the hyperbolic tangent lead to a Schur–Fréchet algorithm for the logarithm that avoids problems associated with the standard "inverse scaling and squaring" method. Inverting the order of evaluation in the logarithmic Fréchet derivative gives a method of evaluating the derivative of the exponential. The resulting Schur–Fréchet algorithm for the exponential gives superior results compared to standard methods on a set of test problems from the literature.

Key words. matrix functions, matrix logarithm, matrix exponential

AMS subject classifications. 15A12, 15A24, 65-04, 65D20, 65F35

PII. S0895479896300334

1. Introduction.

1.1. Background. The discovery of logarithms by John Napier followed a 100-year period in which mathematical notation progressed to the point that algebraic manipulations could be performed with relative ease. Thus the cumbersome logarithmic function defined by Napier was quickly reworked by Henry Briggs into a form that is familiar to us. In 1617, Briggs published a table of logarithms (base 10) followed in 1624 by the more complete *Arithmetica Logarithmica*. The interested reader is referred to Edwards [4] and Goldstine [9] for details. In contrast to his reputation as Napier's drudge, Briggs is viewed by Goldstine as one of the great figures of numerical analysis.

The reason for this high regard is easily seen in the method Briggs used to compute the logarithm of a positive real number. Starting with the fundamental relation $\log(ab) = \log a + \log b$, Briggs wrote $\log a = \log \sqrt{a}\sqrt{a} = 2\log\sqrt{a}$. Repeating this argument gives $\log a = 2^n \log a^{1/2^n}$. Next Briggs noted two facts. First, the repeated roots $a^{1/2^n}$ converge to 1. Second, the logarithm of a number that is very close to 1 can be approximated by using $\log(1+x) \approx cx$, where c is a constant that depends on the base of the logarithm. For Briggs, $c = \log_{10} e \approx 0.4342945$. Using $x = a^{1/2^n} - 1$, we have Briggs' scheme

$$\begin{aligned} \log a &= 2^n \log a^{1/2^n} \\ &\approx 2^n c(a^{1/2^n} - 1). \end{aligned}$$

*Received by the editors March 8, 1996; accepted for publication (in revised form) by B. Kagstrom September 24, 1997; published electronically March 18, 1998. This research was supported in part by Air Force Office of Scientific Research grant F49620-94-1-0104DEF, National Science Foundation grant ECS-9633326, and Office of Naval Research grant N00014-96-1-0456.

http://www.siam.org/journals/simax/19-3/30033.html

†Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 (kenney@seidel.ece.ucsb.edu).

‡University of California, Engineering Dean's Office, One Shields Avenue, Davis, CA 95616-5294 (laub@ucdavis.edu).

So far Briggs is operating at about the level of a modern numerical analyst. Now comes the astounding part. To obtain the accuracy he wanted (about 14 digits), Briggs needed to take up to 54 successive square roots while carrying 32 digits in each computation (i.e., quadruple accuracy on most current computers). To help in this task he invented a finite difference scheme [9] that is equivalent to the binomial expansion of $(1+x)^p$ for the fractional power $p = 1/2$. This was 50 years before Newton stated the binomial theorem in its general form. This procedure was repeated for each of the over 20,000 entries in his table of logarithms (not counting the reduction in effort due to factoring composite numbers). With Goldstine, we can only stand in awe of Briggs' numerical skill and perseverance.

It is helpful to spend a moment looking at the problem of how to select the number of square roots in Briggs' method.

Suppose that $a > 1$. Then $x > 0$ where $x \equiv a^{1/2^n} - 1$. Let x_c denote the value of x that we carry in computation after truncation. That is, we assume $x_c = x - \epsilon$ where $\epsilon \geq 0$ is near the relative machine precision. Here we are ignoring errors in calculating the square roots of a; these errors are generally small in comparison to the error induced by the cancellation of nearly equal terms during the subtraction of 1 from $a^{1/2^n}$.

Use the Taylor series for $\log(1 + x)$ to get

$$\begin{aligned}
\log a &= 2^n \log a^{1/2^n} \\
&= 2^n c \left(x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots \right) \\
&= 2^n c \left((x_c + \epsilon) - \frac{(x_c + \epsilon)^2}{2} + \frac{(x_c + \epsilon)^3}{3} + \cdots \right) \\
&= 2^n c \left(x_c - \frac{x_c^2}{2} + \epsilon(1 - x_c + x_c^2 + \cdots) + \cdots \right) \\
&= 2^n c x_c + 2^n c \left(\frac{\epsilon}{1 + x_c} - \frac{x_c^2}{2} \right) + \cdots
\end{aligned}$$

Thus the relative error in the approximation $\log a \approx 2^n c x_c$ is given (to first order in x_c and ϵ) by

$$\left| \frac{\log a - 2^n c x_c}{2^n c x_c} \right| = \frac{1}{x_c} \left| \frac{\epsilon}{1 + x_c} - \frac{x_c^2}{2} \right|$$

and is approximately minimized at $x_c \approx \sqrt{2\epsilon}$. To see this, note that the expression inside the absolute values is approximately zero when $\frac{x_c^2}{2} \approx \epsilon$. Combining this with $x_c \approx a^{1/2^n} - 1$ tells us approximately how many square roots to take to minimize the error in Briggs' method

$$n \approx \log \left(\frac{\log a}{\sqrt{2\epsilon}} \right) / \log 2.$$

To illustrate, in carrying 32 digits we might expect an error on the order of $\epsilon = 5 \times 10^{-33}$. If $a = 2$ then we want $n \approx 53$ square roots. (For this problem Briggs used 54 square roots.)

Aside from helping us understand Briggs' method this analysis brings out an important point. The number of square roots must be sufficient to ensure the accuracy

of the approximation but not so many as to encounter loss of accuracy in the subtraction of nearly equal terms in forming $x = a^{1/2^n} - 1$. This interpretation also aids us in understanding a conclusion of Dieci, Morini, and Papini [6] concerning square root methods of evaluating the logarithm of a matrix: If an upper triangular matrix A has one subblock A_{11} near the identity while A as a whole is far from the identity, then too many square roots need to be applied to A_{11} to bring A near the identity.

This type of problem can occur in evaluating other matrix functions. For example, the scaling and squaring method [30] for the matrix exponential can result in overscaling for subblocks that are near zero. This paper presents a method of evaluating matrix functions that treats subblocks individually and thus avoids this type of problem.

1.2. Computing functions of matrices by the Schur–Fréchet method. Before considering the logarithmic and exponential functions, it is helpful to look at general functions of matrices. For simplicity we restrict our attention to functions $F = F(A)$ that can be expressed in a convergent power series in a given matrix A or as a power series in some simple transformation of A such as $Y = I - A$. For example,

$$\begin{aligned} e^A &= I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots, \\ A^{1/2} &= I - \frac{1}{2}Y - \frac{1}{8}Y^2 - \frac{1}{16}Y^3 + \cdots, \\ \log A &= -Y - \frac{1}{2}Y^2 - \frac{1}{3}Y^3 - \cdots, \end{aligned}$$

where we assume that $\|Y\| < 1$. Here and in what follows we assume log is the natural logarithm.

In this subsection we discuss evaluating matrix functions by the "Schur–Fréchet" method in which the matrix A is first put in upper triangular form. The main diagonal entries of $F(A)$ are then given by the scalars $F(a_{ii})$ and the rest of the upper diagonals of $F(A)$ can be computed using the Fréchet derivative of F. This approach was first used by Parlett [28], who derived a general finite difference form for the Fréchet derivative based on the fact that A and $F(A)$ commute. Unfortunately, this finite difference approach breaks down if A has multiple eigenvalues and may give inaccurate results if the eigenvalues of A are nearly equal. One way of trying to deal with this difficulty is to reduce A only to a "block triangular" form with the diagonal blocks containing clusters of nearby or identical eigenvalues. Specialized methods for the functions of the diagonal blocks are then used together with appropriate block recurrence formulas. For details on this and related methods, see [15].

The breakdown of Parlett's method, however, is due to the way in which the Fréchet derivative is computed. For example, by using an alternative method of evaluating the Fréchet derivative, Björck and Hammarling [2] developed a stable Schur–Fréchet algorithm for the square root function. The difference between the two methods can be seen in the equality

$$\frac{\sqrt{a_{22}} - \sqrt{a_{11}}}{a_{22} - a_{11}} = \frac{1}{\sqrt{a_{22}} + \sqrt{a_{11}}}.$$

For a_{11} equal to or nearly equal to a_{22}, we may not be able to evaluate the left-hand side accurately but the right-hand side does not pose any problem. Thus the accuracy of the Schur–Fréchet method depends on how the Fréchet derivative is evaluated.

A major contribution of this paper is the derivation of a procedure for evaluating the logarithmic Fréchet derivative that avoids the cancellation effects associated with using the intermediate matrix $A^{1/2^n} - I$. Instead, a hyperbolic form of the logarithmic Fréchet derivative is approximated efficiently and accurately via a rational function in $A^{1/2^n}$; this in turn gives a Schur–Fréchet method for computing the logarithm. The same approximation procedure can be reversed to give the Fréchet derivative of the exponential function; this in turn leads to a Schur–Fréchet algorithm for the exponential of a matrix that avoids some of the "hump" problems encountered by the standard "scaling and squaring" methods of evaluating the matrix exponential [26], [30] (see Example 1).

The Fréchet derivative $L_F(Z, A)$ of F at A in the matrix direction Z is defined by the limit of the Newton quotient for F

$$L_F(Z, A) \equiv \lim_{\delta \to 0} \frac{F(A + \delta Z) - F(A)}{\delta}.$$

The squaring function provides a useful illustration. For notational convenience, let X replace A and set $F(X) = X^2$. The Newton quotient for the squaring function is

$$\frac{(X + \delta Z)^2 - X^2}{\delta} = XZ + ZX + \delta Z^2.$$

Letting $\delta \to 0$ gives the Fréchet derivative of the squaring function at X in the direction Z

$$L_2(Z, X) = XZ + ZX.$$

(Here the subscript "2" denotes the squaring function.) The Fréchet derivatives of a function F and the inverse function F^{-1} are related via the maxim "the Fréchet derivative of the inverse is the inverse of the Fréchet derivative." Thus we find that for the square root function $F(A) = A^{1/2} \equiv X$ the Fréchet derivative at A in the direction Z is given by $L = L_{1/2}(Z, A)$ where L satisfies the Sylvester equation

$$Z = XL + LX.$$

See [2], [16], [24] for more details. See also the related work in [23] and [25].

The following lemma is the basis of the Schur–Fréchet method.

LEMMA 1.1 (Schur–Fréchet). *Let $A = D + Z$, where D and Z have the same block structure*

$$D = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \quad \text{and} \quad Z = \begin{bmatrix} 0 & A_{12} \\ 0 & 0 \end{bmatrix}.$$

Then

$$F(A) = F(D) + L_F(Z, D). \tag{1}$$

Proof. The absence of second-order terms in Z on the right-hand side of (1) is due to the fact that Z is nilpotent of order 2, i.e., $Z^2 = 0$. Moreover, if M is (block) upper triangular then MZ and ZM are also nilpotent of order 2. This means that any second-order or higher terms in Z do not appear in the power series expansion of $F(D + Z)$. See [16] for details on the power series expansion of $F(D + Z)$. □

This result can also be proved by the method of Parlett in [28]. The function $F(A)$ commutes with A; this yields

$$A_{11}F_{12} - F_{12}A_{22} = F_{11}A_{12} - A_{12}F_{22} \tag{2}$$

where F_{ij} is the (i,j) block of $F(A)$. If A has distinct eigenvalues this linear relation can be solved for F_{12}. Thus Parlett's method can be interpreted as a way of computing the Fréchet derivative of F for block strictly upper triangular matrix directions

$$L_F\left(\begin{bmatrix} 0 & A_{12} \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}\right) = \begin{bmatrix} 0 & F_{12} \\ 0 & 0 \end{bmatrix}.$$

For the square root function $F(A) = A^{1/2} = X$, Parlett's method yields

$$A_{11}X_{12} - X_{12}A_{22} = X_{11}A_{12} - A_{12}X_{22}. \tag{3}$$

In contrast to this, Björck and Hammarling [2] show directly that the relation $X^2 = A$ gives

$$X_{11}X_{12} + X_{12}X_{22} = A_{12}. \tag{4}$$

1.3. Logarithms of matrices. The need to compute the logarithm of a matrix occurs in many areas of engineering. As an illustration, suppose that we are studying a system governed by a linear differential equation of the form $dy/dt = Xy$ where the coefficient matrix X is unknown.

Can we recover X from observations of the state vector y? That is, suppose we know the state vector at times $t = 0, h, 2h, \ldots, nh$, where n is the order of the system and h is the sampling time. Let $y_k = y(kh)$. The transition between sampling times is governed by $y_{k+1} = e^{hX}y_k$. This can be written as $Y_1 = e^{hX}Y_0$, where $Y_0 \equiv [y_0, y_1, \ldots, y_{n-1}]$ and $Y_1 \equiv [y_1, y_2, \ldots, y_n]$. From this we see that $e^{hX} = Y_1Y_0^{-1}$. If we can find the logarithm of e^{hX} then we can recover X.

See also Sinha and Lastman [21], [29] for control theory applications involving the transformation from a discrete-time system model to a continuous-time system model. Unfortunately, the iterative procedures for finding the logarithm of a matrix given in [21] and [29] do not always converge even when the logarithm is well defined.

The need to find logarithms of matrices also arises in linear systems with periodic coefficients. Let $dy/dt = A(t)y$, where $A(t+\omega) = A(t)$ for some period ω. In this case, a fundamental matrix solution Φ satisfies $\Phi(t+\omega) = \Phi(t)C$ for some nonsingular matrix C. Let $R = \log(C)/\omega$. Then [3, p. 96, Theorem 2.12] there is a periodic nonsingular matrix $P = P(t)$ of period ω such that $\Phi(t) = P(t)e^{tR}$. Thus, for example, all solutions of $dy/dt = A(t)y$ decay asymptotically as $t \to +\infty$ if the logarithm of the matrix $\Phi^{-1}(0)\Phi(\omega)$ has all of its eigenvalues in the left half-plane.

1.4. Defining the logarithm of a matrix. When is the logarithm of a matrix well defined? In the scalar case, the logarithm x of a number a satisfies $a = e^x$. On the other hand, $x + 2k\pi i$ is also a logarithm of a for any integer k since $e^{x+2k\pi i} = e^x e^{2k\pi i} = e^x$. This nonuniqueness problem can be avoided by using the principal branch of the logarithm. That is, we require x to satisfy $a = e^x$ and restrict the imaginary part of x to lie between $-\pi$ and π. This restriction means that we must exclude values of a on the negative real axis.

For the matrix case the same approach can be used. If A is a matrix with no eigenvalues on the (closed) negative real axis $\mathbb{R}^-$, then there is a unique matrix X

satisfying $A = e^X$ subject to the restriction that the eigenvalues λ of X lie in the infinite strip $-\pi < \mathrm{Im}(\lambda) < \pi$; see [16] and also [5], [8], [12], [14], and [31]. Denoting this matrix X by the symbol $\log A$ we see that out of all matrices M satisfying $A = e^M$, the matrix X is the one with eigenvalues closest to the origin. This is the logarithm that is computed by the algorithms given in this paper.

Additionally, this choice of the logarithm satisfies the equation $X = \frac{1}{p}\log(A^p)$ for $0 < p < 1$. Note, however, that this formula may not hold for $p > 1$ since this may force the eigenvalues of A across the negative real axis. For example, in the scalar case if $a = i$ then $x = i\pi/2$. If we take $p = 4$ then $a^4 = 1$ so $\frac{1}{4}\log a^4 = \frac{1}{4}\log 1 = 0 \neq i\pi/2$.

Example 1. Let us illustrate the above discussion with a simple example that has the further advantage of shedding some light on the "hump" problem of the matrix exponential. Consider

$$X = \begin{bmatrix} \alpha + \beta i & x \\ 0 & \alpha - \beta i \end{bmatrix} \quad \text{and} \quad e^{tX} = \begin{bmatrix} e^{(\alpha+\beta i)t} & xe^{\alpha t}\sin(\beta t)/\beta \\ 0 & e^{(\alpha-\beta i)t} \end{bmatrix}.$$

If $\beta \neq 0$, then for $t = \pi/\beta$ the matrix e^{tX} has a double eigenvalue at $-e^{\alpha t}$ on the negative real axis. At the same time $\sin(\beta t) = 0$ and the (1,2) entry of e^{tX} is zero. Thus at $t = \pi/\beta$ all information about x is lost and we cannot hope to recover X. This phenomenon is distinct from the loss of phase information in the scalar case and shows up as a decrease in the norm of e^{tX}.

This hump problem is discussed by Moler and Van Loan [26] for the squaring phase of the "scaling and squaring" method of evaluating the exponential. In this method, the exponential of $X/2^k$ is approximated (usually by a rational function) where k is taken large enough so that the error in the approximation is negligible. The result is then squared k times to yield the exponential of X. A decrease in norm resulting from the near cancellation of large terms during a squaring has the potential of introducing a loss of accuracy even if the large terms are themselves relatively accurate.

2. The exponential and logarithmic Fréchet derivatives. The difference between the exponentials of $X + L$ and X can be expressed by the standard integral formula

$$e^{X+L} - e^X = \int_0^1 e^{(1-s)X} L e^{s(X+L)}\, ds.$$

Dropping second-order terms in L gives the Fréchet relation

$$Z = \int_0^1 e^{(1-s)X} L e^{sX}\, ds, \tag{5}$$

where Z is the Fréchet derivative of the exponential map at X in the direction L. Alternatively, because of the inverse relationship of the exponential and logarithm functions, L is the Fréchet derivative of the logarithmic map at A in the direction Z where $A = e^X$. See [16] for more details.

The integral expression relating Z and L is difficult to work with numerically. The goal of this section is to reduce this integral expression to a series of coupled equations that are more tractable. The main tool in this reduction is the chain rule for Fréchet derivatives: "the Fréchet derivative of the composition is the composition of the Fréchet derivatives." That is, if $F(X) = G(H(X))$ then $L_F = L_G \circ L_H$ and the

Fréchet relation $Z = L_F(L, X)$ for F can be written as the coupled Fréchet relations of G and H

$$\begin{aligned} Z &= L_G(L_1, H(X)), \\ L_1 &= L_H(L, X). \end{aligned}$$

This can be applied to the exponential map by noting that $e^X = \left(e^{X/2}\right)^2$. That is, the exponential is the composition of a division $X \to X/2$, followed by exponentiation and then squaring. Thus, (5) can be replaced by

$$\begin{aligned} Z &= e^{X/2}L_1 + L_1 e^{X/2} && \text{squaring derivative,} \\ L_1 &= \int_0^1 e^{(1-s)X/2} L_0 e^{sX/2}\, ds && \text{exponential derivative,} \\ L_0 &= L/2 && \text{division derivative.} \end{aligned}$$

An extended version of this can be obtained by using $e^X = (e^{X/2^k})^{2^k}$. This is the composition of a division $X \to X/2^k$, followed by exponentiation and then k squarings. This gives

$$\begin{aligned} Z &= e^{X/2}L_k + L_k e^{X/2} && k\text{th squaring derivative,} \\ &\vdots \\ L_{j+1} &= e^{X/2^{k+1-j}} L_j + L_j e^{X/2^{k+1-j}} && j\text{th squaring derivative,} \\ &\vdots \\ L_2 &= e^{X/2^k} L_1 + L_1 e^{X/2^k} && \text{first squaring derivative,} \\ L_1 &= \int_0^1 e^{(1-s)X/2^k} L_0 e^{sX/2^k}\, ds && \text{exponential derivative,} \\ L_0 &= L/2^k && \text{division derivative.} \end{aligned}$$

In the remainder of this section we concentrate on the logarithmic Fréchet derivative. The results obtained can then be easily converted to results for the exponential Fréchet derivative by using the previously mentioned maxim concerning the relationship between derivatives of inverse functions.

2.1. The hyperbolic tangent connection. The chain rule decomposition expresses the Fréchet derivative of the logarithm at A as the composition of the Fréchet derivative of the logarithm at $A^{1/2^k}$ with a sequence of square root derivatives. The square root derivatives can be evaluated as a series of coupled Sylvester equations. Thus the problem of how to evaluate $L_{\log}(Z, A)$ for a given matrix Z has been reduced to evaluating $L_{\log}(W, A^{1/2^k})$, where W results from applying the Sylvester cascade to Z.

Since we have replaced one logarithmic derivative with another, it might seem that we are no closer to simplifying the Fréchet relation (5). This is not the case, however, if k has been chosen so that $A^{1/2^k}$ is close to the identity: say, $\|I - A^{1/2^k}\| < 0.22$. (See Remark 3 for the rationale behind this choice of the bounding constant.) This means that $L_{\log}(W, A^{1/2^k})$ may be approximated with very low error ($< 10^{-16}$) by using rational functions associated with the hyperbolic tangent function.

This type of approximation is based on a Kronecker form of the logarithmic Fréchet derivative. This form of the Fréchet derivative is derived in the next section in a general setting that can then be applied to the problem of evaluating $L_{\log}(W, A^{1/2^k})$.

2.2. Kronecker form of the Fréchet derivative. The Fréchet derivative L of the logarithm and the Fréchet derivative Z of the exponential at $\mathcal{X}$ are related by the integral expression

$$Z = \int_0^1 e^{(1-s)\mathcal{X}} L e^{s\mathcal{X}}\, ds. \tag{6}$$

(Note: for notational simplicity in this subsection we work with $\mathcal{X}$ rather than $X/2^k$; see the discussion after Lemma 2.2.) It is convenient to rewrite this expression as

$$e^{-\mathcal{X}} Z = \int_0^1 e^{-s\mathcal{X}} L e^{s\mathcal{X}}\, ds. \tag{7}$$

Our basic problem is how to solve for L if we know Z.

Converting this equation to Kronecker form reveals a close connection between Z and L and the hyperbolic tangent function. The Kronecker form of a matrix M is obtained by stacking the columns of M to form a vector $v = \operatorname{vec} M$. We need the following properties [11]:

$$\begin{aligned} A \otimes B &\equiv (a_{ij}B) \qquad (a_{ij} = ij\text{th element of } A),\\ \operatorname{vec}(ABC) &= (C^{\mathrm{T}} \otimes A) \operatorname{vec} B,\\ (A \otimes B)(C \otimes D) &= (AC) \otimes (BD),\\ A \oplus B &\equiv A \otimes I + I \otimes B,\\ e^A \otimes e^B &= e^{A \oplus B}. \end{aligned}$$

Remark 1. Our definition of Kronecker sum is consistent with that of Graham [11] since that definition yields the exponential property as listed. Many authors define the Kronecker sum $A \oplus B$ to be $I \otimes A + B \otimes I$, in which case the exponential property becomes instead $e^A \otimes e^B = e^{B \oplus A}$.

Since the vec operator is linear and since the integral expression on the right-hand side of (7) can be written as the limit of Riemann sums, we may interchange the order of these operators as follows:

$$\begin{aligned} \operatorname{vec} \int_0^1 e^{-s\mathcal{X}} L e^{s\mathcal{X}}\, ds &= \int_0^1 \operatorname{vec}\left(e^{-s\mathcal{X}} L e^{s\mathcal{X}}\right) ds\\ &= \int_0^1 \left(e^{s\mathcal{X}^{\mathrm{T}}} \otimes e^{-s\mathcal{X}}\right) \operatorname{vec} L\, ds\\ &= \int_0^1 e^{s\left(\mathcal{X}^{\mathrm{T}} \oplus (-\mathcal{X})\right)} \operatorname{vec} L\, ds\\ &= \int_0^1 e^{sY}\, ds \operatorname{vec} L \end{aligned}$$

where $Y \equiv \mathcal{X}^{\mathrm{T}} \oplus (-\mathcal{X})$.

Although Y is singular, it is now helpful to work by analogy with the nonsingular scalar case. For $y \neq 0$

$$\begin{aligned} \int_0^1 e^{ys}\, ds &= \frac{e^y - 1}{y}\\ &= \frac{e^y + 1}{2}\ \frac{\tanh(y/2)}{y/2}, \end{aligned}$$

where

$$\tanh(x) \equiv \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

It is convenient to introduce the function $\tau(x) = \tanh(x)/x$. To avoid difficulties with $x = 0$, we define $\tau(x)$ in terms of the associated Taylor series (see [10, p. 35, section 1.411, number 6]),

$$\tau(x) = 1 - \frac{x^2}{3} + \frac{2x^4}{15} - \frac{17x^6}{315} + \cdots. \tag{8}$$

This series is absolutely convergent for $|x| < \pi/2$.

We can combine the above results for $Z = \int_0^1 e^{(1-s)\mathcal{X}} L e^{s\mathcal{X}}\, ds$ by writing

$$\begin{aligned}
\operatorname{vec} Z &= \left(I \otimes e^{\mathcal{X}}\right)\left(I \otimes e^{-\mathcal{X}}\right) \operatorname{vec} Z \\
&= \left(I \otimes e^{\mathcal{X}}\right) \operatorname{vec}\left(e^{-\mathcal{X}} Z\right) \\
&= \left(I \otimes e^{\mathcal{X}}\right) \frac{e^Y + I}{2}\ \tau(Y/2)\ \operatorname{vec} L \\
&= \frac{1}{2}\ \left(e^{\mathcal{X}^{\mathrm{T}}} \otimes I + I \otimes e^{\mathcal{X}}\right) \tau(Y/2)\ \operatorname{vec} L,
\end{aligned}$$

where $\tau(Y/2)$ is defined by the obvious matrix analogue of the power series (8).

Noting that

$$\left(e^{\mathcal{X}^{\mathrm{T}}} \otimes I + I \otimes e^{\mathcal{X}}\right) \operatorname{vec} M = \operatorname{vec}\left(e^{\mathcal{X}} M + M e^{\mathcal{X}}\right),$$

we have the following result.

THEOREM 2.1. *Let $Z = \int_0^1 e^{(1-s)\mathcal{X}} L e^{s\mathcal{X}}\, ds$ and $Y \equiv \mathcal{X}^{\mathrm{T}} \oplus (-\mathcal{X})$, where $\|\mathcal{X}\|_F < \pi/2$. Then Z and L are related by the coupled equations*

$$\begin{aligned}
2Z &= e^{\mathcal{X}} L_0 + L_0 e^{\mathcal{X}}, \\
\operatorname{vec} L_0 &= \ \tau(Y/2)\ \operatorname{vec} L.
\end{aligned}$$

Proof. In view of the above development we need only show that $\|Y/2\|_2 < \pi/2$, which ensures that the series for $\tau(Y/2)$ is well defined. For any vector $v = \operatorname{vec} V$ we have

$$Yv = \operatorname{vec}(V\mathcal{X} - \mathcal{X}V).$$

Now use $\|M\|_F = \|\operatorname{vec} M\|_2$ to get $\|Yv\|_2 \le 2\|\mathcal{X}\|_F\ \|V\|_F = 2\|\mathcal{X}\|_F\ \|v\|_2$. This shows that $\|Y/2\|_2 \le \|\mathcal{X}\|_F < \pi/2$. □

2.3. Rational approximations of $\tau(x) = \tanh(x)/x$. Suppose for a moment that we can replace $\tau(Y/2)$ in Theorem 2.1 by a factored rational function R of Y, say

$$\begin{aligned}
\operatorname{vec}(L_0) &= R(Y)\ \operatorname{vec}(L) \\
&= \prod_{i=1}^{m} (I - Y/\beta_i)^{-1}(I - Y/\alpha_i)\ \operatorname{vec}(L),
\end{aligned}$$

where $\alpha_1, \alpha_2, \ldots, \alpha_m$ are the zeros of R and $\beta_1, \beta_2, \ldots, \beta_m$ are the poles of R.

Since $Y = \mathcal{X}^{\mathrm{T}} \oplus (-\mathcal{X}) = \mathcal{X}^{\mathrm{T}} \otimes I - I \otimes \mathcal{X}$, we have $Y \operatorname{vec} M = \operatorname{vec}(M\mathcal{X} - \mathcal{X}M)$ for any compatibly dimensioned matrix M. Then the preceding rational expression can be written as

$$\begin{aligned}
G_0 &= L,\\
(I/2 + \mathcal{X}/\beta_1)G_1 + G_1(I/2 - \mathcal{X}/\beta_1) &= (I/2 + \mathcal{X}/\alpha_1)G_0 + G_0(I/2 - \mathcal{X}/\alpha_1),\\
&\vdots\\
(I/2 + \mathcal{X}/\beta_m)G_m + G_m(I/2 - \mathcal{X}/\beta_m) &= (I/2 + \mathcal{X}/\alpha_m)G_{m-1} + G_{m-1}(I/2 - \mathcal{X}/\alpha_m),\\
L_0 &= G_m.
\end{aligned}$$

This gives us the final connection between Z and L in the exponential Fréchet relation (5). First we had the Sylvester cascade corresponding to the repeated squaring operations followed by the hyperbolic tangent equation. The above equations show that if the hyperbolic tangent function is approximated by a rational function the result is simply another Sylvester cascade. If the original matrix $A = e^X$ has been put in upper triangular form, the solutions to these Sylvester equations can be found efficiently by standard procedures. Thus the main difficulty remaining is to find rational approximations for the hyperbolic tangent and to bound the error incurred in using these approximations.

The principal Padé approximants of the function $\tanh(x)/x$ can be recovered from the continued fraction expansion [1]

$$\frac{\tanh(x)}{x} = \cfrac{1}{1 + \cfrac{x^2/1\cdot 3}{1 + \cfrac{x^2/3\cdot 5}{1 + \cdots \cfrac{x^2/(2k-1)\cdot(2k+1)}{1+\cdots}}}}$$

For example, the [8,8] Padé approximation to $\tanh(x)/x$ is given by

$$R_8 = \frac{34459425 + 4729725x^2 + 135135x^4 + 990x^6 + x^8}{34459425 + 16216200x^2 + 945945x^4 + 13860x^6 + 45x^8}.$$

Using the theory of orthogonal polynomials and their connection with continued fractions [1], [22] we can show that the zeros and poles of these approximants lie on the imaginary axis.

For comparison we note the Eulerian expansion

$$\frac{\tanh(x)}{x} = \frac{\pi^2 + x^2}{\pi^2 + 4x^2}\ \frac{\pi^2 + x^2/4}{\pi^2 + 4x^2/9} \cdots \frac{\pi^2 + x^2/k^2}{\pi^2 + 4x^2/(2k-1)^2} \cdots, \tag{9}$$

which shows clearly the zeros $(\pm k\pi i)$ and poles $(\pm((2k-1)\pi i/2)$ of $\tanh(x)/x$. See Table 1 for a comparison with the poles and zeros of R_8.

Remark 2. Set the rational expansion (9) equal to the Taylor series (8), multiply both sides by the denominator of the rational expansion, and then equate the coefficients of like powers of x. This is essentially the procedure Euler used [9] to find power series expressions for the even powers of π. For example, the series for π^2 is given by the well-known formula

$$\frac{\pi^2}{6} = \sum_{n=1}^{+\infty} \frac{1}{n^2}.$$

TABLE 1
Eight-digit values for the poles and zeros of R_8 and $\tanh(x)/x$.

	Zeros	Poles
R_8	$\pm 3.1415927\,i$	$\pm 1.5707963\,i$
	$\pm 6.2899752\,i$	$\pm 4.7124693\,i$
	$\pm 10.281299\,i$	$\pm 7.9752405\,i$
	$\pm 28.893970\,i$	$\pm 14.822981\,i$
$\tanh(x)/x$	$\pm 3.1415927\,i$	$\pm 1.5707963\,i$
	$\pm 6.2831853\,i$	$\pm 4.7123890\,i$
	$\pm 9.4247780\,i$	$\pm 7.8539816\,i$
	$\pm 12.566371\,i$	$\pm 10.999557\,i$

The distribution of the poles and zeros allows us to say something about the conditioning of the Sylvester steps in the rational approximation of $\tanh(x)/x$. In the following lemma, γ plays the role of either a pole or a zero; we may assume that $|\gamma| \geq \pi/2$.

LEMMA 2.2. *Let $Y = \mathcal{X}^{\mathrm{T}} \otimes I - I \otimes \mathcal{X}$. If $2\|\mathcal{X}\|_F/|\gamma| < 1$; then*

$$1 - 2\|\mathcal{X}\|_F/|\gamma| \leq \|I + Y/\gamma\|_2 \leq 1 + 2\|\mathcal{X}\|_F/|\gamma|$$

and

$$(1 + 2\|\mathcal{X}\|_F/|\gamma|)^{-1} \leq \|(I + Y/\gamma)^{-1}\|_2 \leq (1 - 2\|\mathcal{X}\|_F/|\gamma|)^{-1}.$$

Proof. Since $Y \operatorname{vec} V = \operatorname{vec}(V\mathcal{X} - \mathcal{X}V)$ we have $\|Y\|_2 \leq 2\|\mathcal{X}\|_F$. Thus $\|I + Y/\gamma\|_2 \leq 1 + \|Y\|_2/|\gamma| \leq 1 + 2\|\mathcal{X}\|_F/|\gamma|$. The other inequalities can be obtained in a similar manner. □

To illustrate, in the decomposition given in Theorem 2.1 we approximate $\tau(Y/2)$ where $Y = \mathcal{X}^{\mathrm{T}} \otimes I - I \otimes \mathcal{X}$ and $\mathcal{X} = X/2^k$. This is the same as approximating $\tau(Y_1)$ where $Y_1 = \mathcal{X}_1^{\mathrm{T}} \otimes I - I \otimes \mathcal{X}_1$ and $\mathcal{X}_1 = X/2^{k+1}$.

If we assume that k is large enough so that $\|X\|_F/2^k < 0.25$ (which is usual for the log problem; see [16]), then $2\|\mathcal{X}_1\|_F/|\gamma| < \frac{1}{2\pi}$ and $I + Y_1/\gamma$ is well conditioned with respect to inversion since

$$\|I + Y_1/\gamma\|_2 \|(I + Y_1/\gamma)^{-1}\|_2 \leq \frac{1 + \frac{1}{2\pi}}{1 - \frac{1}{2\pi}} < 1.38\ .$$

A similar analysis shows that the product of the condition numbers of each Sylvester step corresponding to R_8 is bounded above by

$$\left(\frac{1 + \frac{1}{2\pi}}{1 - \frac{1}{2\pi}}\right)^2 \left(\frac{1 + \frac{1}{4\pi}}{1 - \frac{1}{4\pi}}\right)^2 \cdots \left(\frac{1 + \frac{1}{16\pi}}{1 - \frac{1}{16\pi}}\right)^2 < 5.68.$$

Thus this approximation does not introduce significant error due to ill conditioning of the Sylvester operations.

We still need to ascertain the accuracy of the Padé approximations of $\tanh(x)/x$. Our main result along these lines is the following theorem.

THEOREM 2.3. *Let $R_k(x)$ be the kth Padé approximant to $\tau(x) \equiv \tanh(x)/x$ and let $\tilde{R}_k(x) \equiv R_k(ix)$ be the kth Padé approximant to $\tilde{\tau}(x) \equiv \tau(ix) = \tan(x)/x$. The function*

$$f(x) \equiv \tilde{\tau}(x) - \tilde{R}_k(x)$$

is an increasing function of x for $0 \le x < \pi/2$. For $\|X\| < \pi/2$ the error in the matrix approximation $R_k(X) \approx \tau(X)$ is bounded by the scalar error

$$\|\tau(X) - R_k(X)\| \le f(\|X\|), \tag{10}$$

where $\|\cdot\|$ is any consistent matrix norm.

Proof. The proof uses standard methods like those in [17]. For details see [20]. □

Using the above theorem we see that if k is large enough so that $\|X\|_F/2^k < 0.25$, then the difference between $\tau(Y/2)$ and $R_8(Y/2)$ is less than 10^{-16} where $Y \equiv \left(X^{\mathrm{T}} \otimes I - I \otimes X\right)/2^k$.

3. Algorithms. In this section we synthesize the analysis of the preceding sections to develop algorithms for evaluating the logarithm and exponential of a matrix.

3.1. A Schur–Fréchet algorithm for the logarithm. Assume that A and X have the block structure

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} X_{11} & X_{12} \\ 0 & X_{22} \end{bmatrix}.$$

Suppose that we know A and have already calculated $X_{11} = \log(A_{11})$ and $X_{22} = \log(A_{22})$ but we don't know X_{12}. Then by Lemma 1.1, if D and Z have the same block structure as A with

$$D = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix} \quad \text{and} \quad Z = \begin{bmatrix} 0 & A_{12} \\ 0 & 0 \end{bmatrix},$$

then X_{12} is just the $(1,2)$ block of the Fréchet derivative of the logarithm of D in the matrix direction Z.

Before describing how to find X_{12}, we must address a small detail: we need to determine k such that $\|X/2^k\|_F < 0.25$. However, we are assuming that we only know the main diagonal blocks of X. The next lemma shows how we can get around this difficulty.

LEMMA 3.1. *Let $X = \log(A)$ and $Y = I - A^{1/2^k}$. Assume that k is large enough so that $\|Y\|_F < 1$. Then $\|X/2^k\|_F < |\log(1 - \|Y\|_F)|$.*

Proof. Take norms in the expression $X/2^k = \frac{1}{2^k}\log(A) = -(Y + Y^2/2 + Y^3/3 + \cdots)$. □

Remark 3. Since $0.22 < 1 - e^{-0.25}$, Lemma 3.1 shows that if $\|I - A^{1/2^k}\|_F < 0.22$, then $\|X/2^k\|_F < 0.25$.

ALGORITHM (SCHUR–FRÉCHET LOGARITHM).
Let k be large enough so that $\|I - A^{1/2^k}\|_F < 0.22$.
Step 1 (square root Sylvester cascade)
a) If $k = 0$, set $L_0 = A_{12}$ and go to Step 2.
b) If $k > 0$, solve in sequence for $L_k,\ L_{k-1}, \ldots, L_0$ in the coupled Sylvester equations

$$\begin{aligned} A_{12} &= A_{11}^{1/2} L_k + L_k A_{22}^{1/2}, \\ L_k &= A_{11}^{1/2^2} L_{k-1} + L_{k-1} A_{22}^{1/2^2}, \\ &\vdots \\ L_2 &= A_{11}^{1/2^k} L_1 + L_1 A_{22}^{1/2^k}, \\ L_0 &= 2^k\ L_1. \end{aligned}$$

Step 2 (hyperbolic tangent Sylvester cascade)
a) Let $\alpha_1, \alpha_2, \ldots, \alpha_8$ be the zeros of $p(x) = 34459425 + 4729725x^2 + 135135x^4 + 990x^6 + x^8$.
b) Let $\beta_1, \beta_2, \ldots, \beta_8$ be the zeros of $q(x) = 34459425 + 16216200x^2 + 945945x^4 + 13860x^6 + 45x^8$.
c) Set $\mathcal{X}_{11} = X_{11}/2^k$ and $\mathcal{X}_{22} = X_{22}/2^k$. Solve in sequence for $G_8, G_7, \ldots, G_0$ and $\tilde{X}_{12}$ in the coupled Sylvester equations

$$\begin{aligned} L_0 &= A_{11}^{1/2^k} G_8 + G_8 A_{22}^{1/2^k}, \\ (I + \mathcal{X}_{11}/\beta_8)G_8 + G_8(I - \mathcal{X}_{22}/\beta_8) &= (I + \mathcal{X}_{11}/\alpha_8)G_7 + G_7(I - \mathcal{X}_{22}/\alpha_8), \\ &\vdots \\ (I + \mathcal{X}_{11}/\beta_1)G_1 + G_1(I - \mathcal{X}_{22}/\beta_1) &= (I + \mathcal{X}_{11}/\alpha_1)G_0 + G_0(I - \mathcal{X}_{22}/\alpha_1), \\ \tilde{X}_{12} &= 2G_0. \end{aligned}$$

MATLAB routines implementing the Schur–Fréchet method for the logarithm can be found in [20]. The significance of the hyperbolic tangent approach to approximating the Fréchet derivative of the matrix logarithm function is that the Padé approximation of the hyperbolic tangent function makes use of the matrix $A^{1/2^n}$ rather than $A^{1/2^n} - I$. This avoids the cancellation of nearly equal terms and the concomitant (but unnecessary) loss of accuracy.

3.2. Comparison with related methods for the logarithm. Other methods have been proposed for computing the logarithm of a matrix. Perhaps the most competitive is an adaptation of Briggs' square root method: take k square roots of A so that $A^{1/2^k}$ is close to the identity. Then set $Y = I - A^{1/2^k}$ and use a $[p/q]$ Padé approximation of $\log(I - Y)$, where $[p/q]$ is the order of the Padé approximation. (Briggs used a $[1/0]$ approximation in preparing the scalar logarithm tables in *Arithmetica Logarithmica*.) Details of the Briggs–Padé method can be found in [16] and [17]. This method is implemented with $[p/q] = [8/8]$ in the MATLAB routine `logBP` given in [20].

The Briggs–Padé method has been criticized by Dieci, Morini, and Papini [6] on the grounds that if A is an upper triangular matrix with one subblock A_{11} near the identity while the rest of A is far from the identity, then treating the matrix as a whole results in too many square roots being applied to A_{11} to bring A near the identity. Subsequently, when forming the difference matrix $Y = I - A^{1/2^k}$, there is an unneccesary loss of accuracy in the Y_{11} block because of the subtraction of nearly equal terms. This effect is seen in the following example.

Example 2. Let

$$X = \begin{bmatrix} \alpha & \beta \\ 0 & \alpha \end{bmatrix}, \qquad A = \begin{bmatrix} e^\alpha & \beta e^\alpha \\ 0 & e^\alpha \end{bmatrix}.$$

If we take $\alpha = 1/10$ and $\beta = 10^6$ then 23 square roots are needed to bring A close to I; for $k = 23$, we have $\|I - A^{1/2^k}\|_F < 0.22$. The approximation X_{bp} computed by the MATLAB routine `logBP` has lost about seven digits of accuracy in its main diagonal entries. The relative error matrix E_{bp} defined by $E_{\mathrm{bp}}(i,j) = |X_{\mathrm{bp}}(i,j) - X(i,j)|/(|X(i,j)| + \epsilon)$, where $\epsilon \approx 2.2 \times 10^{-16}$, is given by

$$E_{\mathrm{bp}} = \begin{bmatrix} 8.9 \times 10^{-9} & 5.8 \times 10^{-16} \\ 0 & 8.9 \times 10^{-9} \end{bmatrix}.$$

This problem inspired the development of the Schur–Fréchet method described in this paper. By building up the logarithm via a recursion on the upper subblocks, this method avoids the problem of treating the matrix as a whole in determining the number of square roots. At the same time a rational approximation of a novel form of the Fréchet derivative of the logarithm is used, which does not require the difference matrix $Y = I - A^{1/2^k}$. The Schur–Fréchet method is implemented in the MATLAB routine `logSF` given in [20]. For the above example, the matrix X_{sf}, computed by `logSF`, has the relative error matrix E_{sf} given by

$$E_{\mathrm{sf}} = \begin{bmatrix} 6.9 \times 10^{-16} & 0 \\ 0 & 6.9 \times 10^{-16} \end{bmatrix}.$$

It is interesting to note that because of the evaluation by subblocks, the number of square roots needed by the Schur–Fréchet method for this problem is much smaller than for the Briggs–Padé method. Here we need take only four square roots to ensure that the main diagonal entries are brought within 0.22 of 1.0.

The Briggs–Padé method and the Schur–Fréchet method require successive square roots of A or its subblocks; to compute these efficiently it is convenient to first put A in upper triangular form. Once this is done the work needed to compute the logarithm is approximately the same for both methods; we omit the details of such a comparison and refer the interested reader to [16].

The desire to avoid square roots and transformations to upper triangular form has prompted some researchers to formulate other methods related to the Taylor series for the logarithm. For example, Dieci, Morini, and Papini [6] note that if A has all of its eigenvalues in the right half-plane, then $C = (I - A)(I + A)^{-1}$ has all of its eigenvalues inside the unit circle. This implies that the following series in C converges: $\log A = \log(I - C) - \log(I + C) = -2(C + C^3/3 + C^5/5 + \cdots)$.

This series for the logarithm was given by Gregory in 1668 in *Exercitationes Geometricae* (see Goldstine [9, pp. 60–61]. The MATLAB routine `logG` given in [20] implements a truncated version of Gregory's series. The analysis of Luke in [22] shows that if we switch from Gregory's series to the principal Padé approximants of $\log\left((I - C)(I + C)^{-1}\right)$, then the requirement that the eigenvalues of A lie in the right half-plane can be relaxed because the main diagonal Padé approximants converge as long as A does not have eigenvalues on $\mathbb{R}^-$.

However, the convergence may be very slow. For example, in Gregory's series for the scalar problem, if $a = 1000$ we find that the first 200 terms of the series give an approximation of $\log(a)$ with a relative error of 0.0446 while 2000 terms are needed to reduce the relative error to 5.4×10^{-6}! This would certainly be unacceptable in the matrix case since each additional term requires another matrix multiplication. More serious for matrix problems is the cancellation of nearly equal terms in the series; this is why the globally convergent Taylor series for the matrix exponential is unsuitable for computational purposes (see [26]). In contrast to this, for $a = 1000$, the Briggs–Padé method needs six square roots to ensure that $a^{1/2^k} - 1 < 0.22$ and then seven multiplications and a division to approximate $\log(a)$ with a relative error of less than 10^{-16}.

Somewhat suprisingly, the Gregory series method performs quite well on Example 2. In part this can be attributed to the fact that the matrix C is nearly nilpotent of order two.

3.3. The exponential of a matrix. Although there are many ways to approximate the exponential of a matrix [26], the most commonly used methods rely on the

"scaling-and-squaring" formula

$$e^X = \left(e^{X/2^k}\right)^{2^k}, \tag{11}$$

where k is some nonnegative integer. Typically, k is taken large enough so that $e^{X/2^k}$ is easily approximated, usually by a rational function [30], which is then squared k times to give an approximation of e^X.

In [27], Najfeld and Havel give an encyclopedic treatment of the Fréchet derivative of the exponential and present a computational procedure for evaluating the matrix exponential by exploiting the relationship $e^{2B} = (H(B) - B)^{-1}(H(B) + B)$ where

$$H(x) \equiv x\coth(x) = x\,\frac{e^{2x}+1}{e^{2x}-1}.$$

The function H can be approximated using rational expressions that are the reciprocals of those we have given for $\tanh(x)/x$. The algorithm given in [27] is similar to the standard scaling and squaring procedure. To approximate e^X, perform the following:

1) Set $B = X/2^{d+1}$ where d is an integer large enough so that $\|B^2\| < \gamma$ where γ is a prescribed constant (see Section 2.3 in [27] and Example 3 below) near 1.
2) Compute a rational Padé approximation $R = R(B)$ of $H(B)$.
3) Set $E = (R-B)^{-1}(R+B)$ and square the result d times. The matrix E^{2^d} is the computed approximation of e^X.

The problem with this approach and the standard scaling and squaring method is that treating the matrix as a whole can result in overscaling some components. For example, if X is in block upper triangular form with the subblock X_{11} small in norm compared to the rest of X, then with the scaling and squaring approach we will see a relative loss of accuracy in the computed value of the (1,1) subblock of e^X. We emphasize that this overscaling problem may not be apparent if it is hidden by a unitary similarity: $\tilde{X} = UXU^H$.

Example 3. To illustrate, radioactive decay problems often give rise to matrices X and their exponentials of the form [7]

$$X = \begin{bmatrix} -1 & b \\ 0 & -b \end{bmatrix}, \qquad e^X = \begin{bmatrix} e^{-1} & b(e^{-1}-e^{-b})/(b-1) \\ 0 & e^{-b} \end{bmatrix},$$

where $b \neq 1$ determines the relative decay rate.

Table 2 gives the relative error in the (1,1) component of the computed result using the MATLAB scaling-and-squaring Padé routine `expm` (column 2) and the scaling-and-squaring hyperbolic cotangent method of Najfeld and Havel (column 3; for this method we used the 8th-order rational approximation H_8 and $\gamma = 1.151922$, as recommended in [27]). The reason for the loss of accuracy described in Table 2 for these scaling-and-squaring methods can be seen in the approximation $e^x \approx 1 + x$. If x is small compared to 1 then finite wordlength representation causes a loss of approximately $-\log_{10}(|x|)$ decimal digits of x in forming $1 + x$ or e^x. Repeated squaring does not restore these lost digits, which may play a significant role in determining the value of $e^{2^k x}$. As an extreme example, if x is less than the relative machine precision, then e^x will be set equal to 1 and repeated squaring leaves this value unchanged. See Kenney and Laub [20] for a related discussion concerning unnecessary loss of accuracy in evaluating the logarithm by the "inverse scaling-and-squaring" method. It is worth remarking that in real decay rate problems the value of b can be as high as 10^{16}.

TABLE 2
Relative error in the $(1,1)$ *component of* e^X *for Example* 3 *using two scaling-and-squaring methods.*

b	Relative error `expm`	Relative error Najfeld and Havel
10^1	10^{-14}	10^{-15}
10^3	10^{-12}	10^{-14}
10^5	10^{-10}	10^{-12}
10^7	10^{-8}	10^{-11}

As a specific example from [7], the decay chain $\mathrm{Kr}^{90} \overset{33\ \mathrm{sec}}{\rightarrow} \mathrm{Rb}^{90} \overset{2.7\ \mathrm{min}}{\rightarrow} \mathrm{Sr}^{90} \overset{28\ \mathrm{yr}}{\rightarrow} \mathrm{Y}^{90} \overset{65\ \mathrm{hr}}{\rightarrow} \mathrm{Zr}^{90}$ (stable) can be represented in matrix form (with the unit of time equal to 28 years) as

$$X = \begin{bmatrix} 0 & 3.8 \times 10^3 & 0 & 0 & 0 \\ 0 & -3.8 \times 10^3 & 1 & 0 & 0 \\ 0 & 0 & -1 & 5.5 \times 10^6 & 0 \\ 0 & 0 & 0 & -5.5 \times 10^6 & 2.7 \times 10^7 \\ 0 & 0 & 0 & 0 & -2.7 \times 10^7 \end{bmatrix}.$$

If we use `expm` to compute e^X then the relative error in the (3,3) entry is on the order of 10^{-11}, indicating that we have lost about five digits of accuracy. Similar problems in evaluating the matrix exponential via the scaling-and-squaring method can occur in control problems with fast and slow system modes.

The Schur–Fréchet method can avoid problems of this type. Suppose that X and $A = e^X$ have the block structure

$$X = \begin{bmatrix} X_{11} & X_{12} \\ 0 & X_{22} \end{bmatrix} \quad \text{and} \quad A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

Assume that we know X and have already calculated $A_{11} = e^{X_{11}}$ and $A_{22} = e^{X_{22}}$ but that we don't know A_{12}. The development of the preceding sections gives the following algorithm for computing $\tilde{A}_{12}$ where, in exact arithmetic, $\|A_{12} - \tilde{A}_{12}\|_F < 10^{-16}$.

ALGORITHM (SCHUR–FRÉCHET EXPONENTIAL).
Given X_{11}, X_{12}, X_{22}, let $k \geq 0$ be large enough so that $\|X_{ii}\|_F \leq 2^{k-2}$ for $i = 1, 2$.

Step 1 (hyperbolic tangent Sylvester cascade)
- (a) Let $\alpha_1, \alpha_2, \ldots, \alpha_8$ be the zeros of $p(x) = 34459425 + 4729725x^2 + 135135x^4 + 990x^6 + x^8$
- (b) Let $\beta_1, \beta_2, \ldots, \beta_8$ be the zeros of $q(x) = 34459425 + 16216200x^2 + 945945x^4 + 13860x^6 + 45x^8$
- (c) $D_{12} \longleftarrow X_{12}$, $M_{11} \longleftarrow X_{11}/2^k$, $M_{22} \longleftarrow X_{22}/2^k$
- (d) `For` $i = 1$ to 8
  $Q \longleftarrow (I/2 + M_{11}/\alpha_i)D_{12} + D_{12}(I/2 - M_{22}/\alpha_i)$
  Solve for D_{12} in $Q = (I/2 + M_{11}/\beta_i)D_{12} + D_{12}(I/2 - M_{22}/\beta_i)$
  `End for`

Step 2 (top down square root Sylvester cascade)
If $k > 0$, then $X_{ii} \longleftarrow X_{ii}/2$ and $A_{ii} \longleftarrow e^{X_{ii}}$ for $i = 1, 2$
`For` $j = 1$ to k
$D_{12} \longleftarrow A_{11}D_{12} + D_{12}A_{22}$

If $j < k$, then $X_{ii} \longleftarrow X_{ii}/2$ and $A_{ii} \longleftarrow e^{X_{ii}}$ for $i = 1, 2$
`End for`
Repeat the last Sylvester step $D_{12} \longleftarrow A_{11}D_{12} + D_{12}A_{22}$

Step 3 Scale the final result: $A_{12} \longleftarrow D_{12}/2^{k+1}$.

This algorithm might be called the completely recursive form of the Schur–Fréchet approach in that it calls for separate calculation of the coefficient matrices $e^{X_{ii}/2^j}$ in the square root Sylvester cascade. While this appears to give the most accurate results it becomes very expensive as the order of the matrix grows. As a compromise between accuracy and efficiency we have modified the above algorithm to a "semirecursive" form: if X_{ii} is not too large in norm (in Step 2) replace the exponentiation $A_{ii} \longleftarrow e^{X_{ii}/2}$ by the analytically equivalent square root operation $A_{ii} \longleftarrow A_{ii}^{1/2}$. Since A_{ii} is in upper triangular form we can form $A_{ii}^{1/2}$ efficiently by using the stable square root algorithm of Björck and Hammarling [2]; for strictly real matrices see the method of Higham in [13].

For the numerical results reported here, we replaced the exponentiation with the square root if the condition $\|X_{ii}\|_F <$ `tol` with `tol`=100 was satisfied. A condition of this type is needed to protect against destructive underflow. For example, in calculating e^{-1000} the result underflows to zero; if we then start taking square roots to approximate $e^{-1000/2^j}$ the resulting sequence remains stuck at 0 even though it should be converging to 1.

The above algorithm can be used to build up the exponential of an upper triangular matrix starting with the smallest 2×2 blocks on the main diagonal and then doubling the block size at each step. (If the matrix order is not a power of 2, then the lower right-hand block size must be adjusted accordingly.) Using this procedure, and evaluating the coefficient matrices $A_{ii} = e^{X_{ii}/2^j}$ by taking square roots, the exponential of an upper triangular matrix can be computed using approximately $(17 + 5k/3)n^3/6$ scalar multiplications where k is large enough so that $\|X\|_F < 2^{k-2}$. This compares with approximately $(8 + k)n^3/6$ scalar multiplications for the scaling-and-squaring Padé approximation method for an upper triangular matrix. Thus, the Schur–Fréchet method requires about twice the effort as the usual scaling-and-squaring method. However, this does not include the cost of the initial transformation to Schur form and the subsequent back transformation which together add approximately $10n^3$ scalar multiplications. When this extra cost is accounted for, the relative difference between the two methods is much smaller. For example, for a nominal value of $k = 6$, the Schur–Fréchet method requires about $14.5n^3$ multiplications versus $12.3n^3$ for the scaling-and-squaring Padé method.

Table 3 shows that for Example 3 the semirecursive and the completely recursive Schur–Fréchet methods give excellent accuracy compared to the standard scaling-and-squaring method. In this table the relative error in each entry is calculated using

$$\mathrm{relerr}(i, j) = \frac{|e(i, j) - a(i, j)|}{\epsilon + |e(i, j)|},$$

where e and a are, respectively, the exact and computed values of e^X and $\epsilon \approx 2.2 \times 10^{-16}$ is the relative machine epsilon for MATLAB on a Sun SPARCstation. The relative error reported in Table 3 is the Frobenius norm of the relative error matrix.

The Schur–Fréchet method also seems to perform well on the hump problem in Example 1.

Table 4 gives the relative error for Example 1 for the parameters $x = 10^2, \alpha = -1$, and $\beta = \pi - \delta$ for different values of δ. As $\delta \to 0$ the scaling-and-squaring method

TABLE 3
Relative error in the computed value of e^X for Example 3.

b	Relative error `expm` (MATLAB)	Relative error Schur–Fréchet semirecursive	Relative error Schur–Fréchet recursive
10^1	10^{-14}	10^{-15}	10^{-15}
10^3	10^{-12}	10^{-15}	10^{-15}
10^5	10^{-10}	10^{-15}	10^{-15}
10^7	10^{-8}	10^{-15}	10^{-15}

TABLE 4
Relative error in the computed value of e^X for Example 1.

δ	Relative error `expm` (MATLAB)	Relative error Schur–Fréchet semirecursive	Relative error Schur–Fréchet recursive
10^{-1}	10^{-13}	10^{-15}	10^{-15}
10^{-3}	10^{-13}	10^{-15}	10^{-15}
10^{-5}	10^{-11}	10^{-15}	10^{-15}
10^{-7}	10^{-9}	10^{-15}	10^{-15}
10^{-9}	10^{-6}	10^{-15}	10^{-15}

loses accuracy.

The semirecursive Schur–Fréchet method has also been applied to the set of problems in [15] and [16] and in all cases tested gave results that were at least as accurate as the scaling-and-squaring method.

We end this section with a short discussion of how the preliminary Schur transformation affects the accuracy of the subsequent evaluation of the exponential by the Schur–Fréchet method. Specifically, can an ill-conditioned eigensystem for X unduly compromise the accuracy of the Schur–Fréchet approximation of e^X? The answer appears to be no; if X has an ill-conditioned eigensystem, then the mapping $X \mapsto e^X$ is inherently sensitive. To see this suppose that X has an eigenvalue $\lambda = \lambda(X)$ that is sensitive in the sense that there exists a matrix E such that

$$\left.\frac{\partial \lambda(X + tE)}{\partial t}\right|_{t=0} \gg 1.$$

Then we have that the relative sensitivity of the corresponding eigenvalue of e^X is also large:

$$\left.\frac{1}{e^\lambda}\frac{\partial e^{\lambda(X+tE)}}{\partial t}\right|_{t=0} = \left.\frac{\partial \lambda(X + tE)}{\partial t}\right|_{t=0} \gg 1.$$

This is illustrated by Example 4 from [15].

Example 4. Let $X = (H_2SH_1)J(H_1S^{-1}H_2)$ where J is all zeros except

$$\operatorname{diag}(J) = [1, 1, 1, 1, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3]$$

and $J_{12} = J_{23} = J_{34} = 1$, $S = \operatorname{diag}(1, \sigma, \sigma^2, \ldots, \sigma^9)$, $H_1 = I - 0.2ee^T$, and $H_2 = I - 0.2ff^T$ with $e = [1, 1, \ldots, 1]^T$ and $f = [1, -1, 1, \ldots, -1]^T$. The parameter σ is used to control the ill conditioning of the eigensystem of X. Table 5 gives the relative error in the computed value of e^X for various values of σ.

TABLE 5
Relative error in the computed value of e^X for Example 4.

σ	Relative error `expm` (MATLAB)	Relative error Schur–Fréchet semirecursive	Relative error Schur–Fréchet recursive	Cond(H_2SH_1)
0.20	10^{-5}	10^{-7}	10^{-7}	2×10^6
0.18	10^{-4}	10^{-5}	10^{-5}	5×10^6
0.16	10^{-1}	10^{-4}	10^{-4}	1×10^7
0.14	10^{0}	10^{-4}	10^{-4}	5×10^7
0.12	10^{2}	10^{-3}	10^{-3}	2×10^8
0.10	10^{8}	10^{-1}	10^{-1}	2×10^9

4. A posteriori condition estimates. A new procedure has recently been developed by the authors [18] for estimating the sensitivity of each entry in a matrix function. To understand this approach suppose that we are interested in the sensitivity of the (i, j) entry of a matrix function $F = F(A)$. This sensitivity can be measured by the norm of the gradient of the map $A \mapsto F_{ij}(A)$. The basic idea in [18] is to check for sensitivity by perturbing the argument (A) in a random fashion and looking at the resulting effect on the function value. In general, a random perturbation is not going to point in the direction of the gradient (i.e., in the matrix direction that produces the maximal perturbation in $F_{ij}(A)$). Because of this, a scaling factor ω_m, also called the "Wallis" factor, must be introduced. The Wallis factor depends only on the number m of arguments being perturbed. In the case considered here, $m = n^2$, where n is the order of the matrix A. For computational purposes the approximation $\omega_m \approx \sqrt{2/(\pi(m - 0.5))}$ is sufficiently accurate; see [18].

The condition estimation procedure introduced in [18] has a particularly simple form when expressed in terms of the Fréchet derivative $L_F(\cdot, A)$ of F:

1) Select $Z \in \mathrm{N}^{n\times n}$, i.e., select Z uniformly and randomly from the space of all unit Frobenius norm matrices of size n by n. This can be done by setting $Z = \tilde{Z}/\|\tilde{Z}\|_F$, where the entries of the matrix $\tilde{Z}$ have been generated independently with a N(0,1) distribution (i.e., normal with mean 0 and variance 1). Set $m = n^2$.
2) Set $D = \frac{1}{\omega_m} L_F(Z, A)$. Then the expected value of $|d_{i,j}|$ is equal to the norm of the gradient of the mapping $A \mapsto F_{ij}(A)$.

Note that we get a condition estimate for all the entries of F with the evaluation of D. For the purposes of condition estimation, the evaluation of $L_F(Z, A)$ can be performed via the approximation

$$L_F(Z, A) \approx \frac{F(A + hZ) - F(A)}{h},$$

where h is a small positive number, say, $h = 10^{-8}$. (See [18] for more discussion of this point.) Because of this, the above condition estimation procedure should not require more than one extra function evaluation beyond that of $F(A)$. Generally, this computational cost can be reduced considerably by working directly with the explicitly known Fréchet derivative for particular functions such as the square root and logarithm.

The accuracy of this method of estimating the norm of the gradient has been analyzed in terms of the beta distribution [18]. This analysis shows that the procedure gives a first-order estimate of the norm of the gradient, i.e., the probability that the estimate is off by a factor of 10 is about 1/10, the chance of being off by a factor

of 100 is about 1/100, etc. Higher-order estimates are easily generated. A pth-order estimate takes about p times the computational effort of a first-order estimate and the chance of its being off by a factor γ is proportional to $1/\gamma^p$ [18].

Applying this to the square root problem with $p = 1$ gives the following procedure.

PROCEDURE 1 (square root condition estimation).

1) Select $Z \in \mathrm{N}^{n\times n}$.
2) Solve for L in $Z = XL + LX$ and define $D = |L|/\omega_m$, where $m = n^2$ and $|L|$ denotes the matrix with entries $|\ell_{ij}|$.

Example 5. As an illustration, let

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

Then the exact matrix of gradient norms is given by

$$G = \begin{bmatrix} \frac{\sqrt{4+x^2}}{4} & \frac{\sqrt{4+2x^2+x^4}}{4} \\ \frac{1}{2} & \frac{\sqrt{4+x^2}}{4} \end{bmatrix}.$$

If we let $x = 1000$, then to four decimal places

$$G = \begin{bmatrix} 250 & 2.5 \times 10^5 \\ 0.50 & 250 \end{bmatrix}.$$

Using the MATLAB commands

```
Z=randn(2,2);
Z=Z/norm(Z,'fro');
```

we generated

$$Z = \begin{bmatrix} 0.4688 & 0.3338 \\ -0.2187 & -0.7880 \end{bmatrix}.$$

Solving for L in $Z = XL + LX$ and setting $D = |L|/\omega_4$ gave

$$D = \begin{bmatrix} 1.294 \times 10^2 & 1.286 \times 10^5 \\ 2.576 \times 10^{-1} & 1.279 \times 10^2 \end{bmatrix},$$

which, considering that condition numbers need only be accurate to within an order of magnitude, compares very well with the exact matrix G of condition numbers. Of course, this is just one random sample, but the quality of the condition estimates reflects the predictions of the work in [18]. For example, in a second random sample we obtained

$$Z = \begin{bmatrix} 0.4725 & -0.3331 \\ -0.7415 & 0.3405 \end{bmatrix}$$

and

$$D = \begin{bmatrix} 4.373 \times 10^2 & 4.373 \times 10^5 \\ 8.736 \times 10^{-1} & 4.372 \times 10^2 \end{bmatrix},$$

which also compares well with G.

The small-sample statistical condition estimation method can also be applied to the logarithmic map $A \mapsto X = \log A$. This is summarized in the following procedure.

PROCEDURE 2 (logarithm condition estimation).

1) Select $Z \in \mathrm{N}^{n\times n}$.
2) Solve for L in $Z = \int_0^1 e^{(1-s)X} L e^{sX} ds$. This can be done by using the Schur–Fréchet logarithm algorithm or by using the Newton quotient approximation $L \approx (\log(A + hZ) - \log(A))/h$, where h is a small positive number (say, $h = 10^{-8}$). After solving for L, define $D = |L|/\omega_m$ where $m = n^2$ and $|L|$ denotes the matrix with entries $|\ell_{ij}|$.

Example 6. Let T be a nonsingular matrix and define $A = T\tilde{A}T^{-1}$, where $\tilde{A}$ is given by

$$\tilde{A} = \begin{bmatrix} e^{\alpha} & 0 & 0 & 0 \\ 0 & e^{-\alpha} & 0 & 0 \\ 0 & 0 & \cos\theta & -\sin\theta \\ 0 & 0 & \sin\theta & \cos\theta \end{bmatrix}.$$

The parameters α and θ allow us to generate ill-conditioned matrix logarithm problems; these occur as α gets large or as $\theta \to \pi$. The logarithm of A is given by $X = T\tilde{X}T^{-1}$ where

$$\tilde{X} = \begin{bmatrix} \alpha & 0 & 0 & 0 \\ 0 & -\alpha & 0 & 0 \\ 0 & 0 & 0 & -\theta \\ 0 & 0 & \theta & 0 \end{bmatrix}.$$

To illustrate, let T be

$$T = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \\ 0 & 0 & 9 & 10 \\ 0 & 0 & 11 & 12 \end{bmatrix}.$$

If we let $\alpha = 5$ and $\theta = 3.14$, then Procedure 2 yields the following matrix of condition estimates for the logarithm of A:

$$D = \begin{bmatrix} 7.964\times 10^3 & 1.593\times 10^3 & 4.887\times 10^6 & 4.035\times 10^6 \\ 2.406\times 10^4 & 4.810\times 10^3 & 1.043\times 10^7 & 8.619\times 10^6 \\ 8.735\times 10^2 & 1.739\times 10^2 & 1.311\times 10^7 & 1.083\times 10^7 \\ 1.071\times 10^3 & 2.133\times 10^2 & 1.585\times 10^7 & 1.311\times 10^7 \end{bmatrix}.$$

We can compare this with the matrix of "true" condition estimates G as calculated by using finite difference estimates on each entry of the matrix A (this approach entails n^2 extra function evaluations and hence is generally too expensive for even moderately large values of n where n is the order of A):

$$G = \begin{bmatrix} 1.248\times 10^4 & 2.497\times 10^3 & 1.022\times 10^7 & 8.434\times 10^6 \\ 3.803\times 10^4 & 7.606\times 10^3 & 2.181\times 10^7 & 1.802\times 10^7 \\ 3.459\times 10^3 & 6.887\times 10^2 & 2.754\times 10^7 & 2.276\times 10^7 \\ 4.247\times 10^3 & 8.458\times 10^2 & 3.332\times 10^7 & 2.754\times 10^7 \end{bmatrix}.$$

The entries of D are all within a factor of four of the respective entries of G.

The random perturbation method of condition estimation is very flexible and can easily be adapted to problems in which the perturbation must preserve some

structure of the matrix such as symmetry. For brevity, we illustrate how to do this for a particular example and refer the interested reader to a related discussion in [19] for the matrix sign function.

Example 7. The matrix A in Example 6 is block upper triangular whenever T is block upper triangular. For any algorithm that respects this upper triangular structure we don't expect to see perturbation effects arising from perturbations in the lower part of A. To estimate the sensitivity of the logarithm of A with respect to perturbations in the block upper entries we can first set up a random matrix $\tilde{Z}$

$$\tilde{Z} = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix}$$

where the starred entries are selected randomly and independently from an N(0,1) distribution. Set $Z = \tilde{Z}/\|\tilde{Z}\|_F$. Now go to step 2 in Procedure 2; use $m = 12$ rather than $m = n^2 = 16$ in calculating ω_m. For the parameter values in Example 6 this gives

$$D = \begin{bmatrix} 1.485 \times 10^3 & 2.969 \times 10^2 & 2.345 \times 10^7 & 1.936 \times 10^7 \\ 4.453 \times 10^3 & 8.905 \times 10^2 & 5.003 \times 10^7 & 4.134 \times 10^7 \\ 0 & 0 & 6.340 \times 10^7 & 5.240 \times 10^7 \\ 0 & 0 & 7.672 \times 10^7 & 6.340 \times 10^7 \end{bmatrix}.$$

This compares well with the matrix G of "true" restricted condition numbers

$$G = \begin{bmatrix} 2.989 \times 10^3 & 5.977 \times 10^2 & 8.908 \times 10^6 & 7.355 \times 10^6 \\ 8.966 \times 10^3 & 1.793 \times 10^3 & 1.901 \times 10^7 & 1.571 \times 10^7 \\ 0 & 0 & 2.408 \times 10^7 & 1.990 \times 10^7 \\ 0 & 0 & 2.914 \times 10^7 & 2.408 \times 10^7 \end{bmatrix}.$$

5. Conclusion. A Schur–Fréchet algorithm has been presented for the computation of the logarithms of matrices with no eigenvalues on the negative real axis. The Schur–Fréchet method given in this paper applies to matrix functions in general and provides a theoretical framework that includes the work of Parlett [28], Björck and Hammarling [2], and others. The accuracy of the Schur–Fréchet method depends on how accurately the Fréchet derivative can be evaluated.

By deriving a new expression for the Fréchet derivative of the logarithm in terms of the function $\tau(x) = \tanh(x)/x$, together with a Padé approximation of τ, we have developed a procedure for evaluating the Fréchet derivative that avoids the cancellation problems encountered in other methods.

This evaluation process is completely reversible and hence provides a method of evaluating the derivative of the matrix exponential. This leads to a Schur–Fréchet method for evaluating the exponential that avoids some overscaling problems associated with the standard scaling and squaring method. In its completely recursive form the Schur–Fréchet algorithm presented here is not computationally efficient; however, this problem can be avoided by using a semirecursive algorithm based on a tolerance parameter. Large values of this tolerance parameter lead to the square root approach of generating the coefficient matrices in the Sylvester cascade and result in a computational cost that is nearly the same as the scaling-and-squaring Padé approximation method. As the tolerance value is decreased, the accuracy of the result improves but

the computational cost also increases. Theoretical aspects of the error propagation in this semirecursive Schur–Fréchet method need further investigation but the gain in computational efficiency does not appear to compromise accuracy compared with the fully recursive Schur–Fréchet method for moderately large values of the tolerance parameter.

REFERENCES

[1] G. A. Baker, *Essentials of Padé Approximants*, Academic Press, New York, 1975.

[2] A. Björck and S. Hammarling, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.

[3] F. Brauer and J. A. Nohel, *The Qualitative Theory of Ordinary Differential Equations*, W.A. Benjamin, New York, 1969.

[4] C. H. Edwards, Jr., *The Historical Development of the Calculus*, Springer-Verlag, New York, 1979.

[5] W. Culver, *On the existence and uniqueness of the real logarithm of a matrix*, in Proc. Amer. Math. Soc., 17 (1966), pp. 1146–1151.

[6] L. Dieci, B. Morini, and A. Papini, *Computational techniques for real logarithms of matrices*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 570–593.

[7] M. Eisenbud, *Environmental Radioactivity*, McGraw–Hill, New York, 1963.

[8] F. R. Gantmacher, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1959.

[9] H. H. Goldstine, *A History of Numerical Analysis from the* 16*th through the* 19*th Century*, Springer-Verlag, New York, 1977.

[10] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 4th ed., Academic Press, New York, 1965.

[11] A. Graham, *Kronecker Products and Matrix Calculus with Applications*, John Wiley, New York, 1981.

[12] B. Helton, *Logarithms of matrices*, Proc. Amer. Math. Soc., 19 (1968), pp. 733–738.

[13] N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.

[14] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, New York, 1991.

[15] B. Kågström, *Numerical Computation of Matrix Functions*, Tech. report UMINF-58.77, Department of Information Processing, University of Umeå, Umeå, Sweden, 1977.

[16] C. S. Kenney and A. J. Laub, *Condition estimates for matrix functions*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 191–209.

[17] C. S. Kenney and A. J. Laub, *Padé error estimates for the logarithm of a matrix*, Internat. J. Control, 50 (1989), pp. 707–730.

[18] C. S. Kenney and A. J. Laub, *Small-sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comp., 15 (1994), pp. 36–61.

[19] C. S. Kenney and A. J. Laub, *The matrix sign function*, IEEE Trans. Automat. Control, 40 (1995), pp. 1330–1348.

[20] C. S. Kenney and A. J. Laub, *A Schur–Fréchet Algorithm for the Logarithm of a Matrix*, Tech. report, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA, 1995.

[21] G. J. Lastman and N. K. Sinha, *Infinite series for logarithm of matrix applied to identification of linear continuous-time multivariable systems from discrete-time models*, Electron. Lett., 27 (1991), pp. 1468–1470.

[22] Y. L. Luke, *The Special Functions and Their Approximations*, Vols. I and II, Academic Press, New York, 1969.

[23] R. Mathias, *Evaluating the Fréchet derivative of the matrix exponential*, Numer. Math., 63 (1992), pp. 213–226.

[24] R. Mathias, *Condition estimation for matrix functions via the Schur decomposition*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 565–578.

[25] R. Mathias, *Approximation of matrix-valued functions*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 1061–1063.

[26] C. B. Moler and C. F. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836.

[27] I. Najfeld and T. F. Havel, *Derivatives of the matrix exponential and their computation*, Adv. Appl. Math., 16 (1995), pp. 321–375.

[28] B. N. Parlett, *A recurrence among the elements of functions of triangular matrices*, Linear Algebra Appl., 14 (1976), pp. 117–121.
[29] N. K. Sinha and G. J. Lastman, *Transformation algorithm for identification of continuous-time multivariable systems from discrete data*, Electron. Lett., 17 (1981), pp. 779–780.
[30] R. C. Ward, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1977), pp. 600–614.
[31] 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. 19, No. 3, pp. 664–681, July 1998

004

CHOOSING POLES SO THAT THE SINGLE-INPUT POLE PLACEMENT PROBLEM IS WELL CONDITIONED*

VOLKER MEHRMANN† AND HONGGUO XU‡

Abstract. We discuss the single-input pole placement problem (SIPP) and analyze how the conditioning of the problem can be estimated and improved if the poles are allowed to vary in specific regions in the complex plane. Under certain assumptions we give formulas as well as bounds for the norm of the feedback gain and the condition number of the closed loop matrix. Via several numerical examples we demonstrate how these results can be used to estimate the condition number of a given SIPP problem and also demonstrate how to select the poles to improve the conditioning.

Key words. pole placement, condition number, perturbation theory, Jordan form, Cauchy matrix, stabilization, feedback gain, distance to uncontrollability, optimal conditioning

AMS subject classifications. 65F15, 65F35, 65G05, 93B05, 93B55

PII. S0895479896302382

1. Introduction.

We consider linear single-input control systems

$$\dot{x} := dx/dt = Ax + bu, \qquad x(0) = x_0, \tag{1}$$

where $A \in \mathcal{C}^{n\times n}$, $b \in \mathcal{C}^n$, x, u are functions defined on $[0, +\infty) \to \mathcal{C}^n$ and $[0, +\infty) \to \mathcal{C}$, respectively. For such systems, we consider the single-input pole placement problem.

Single-input pole placement (SIPP). *For a given set of poles $\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\}$, find a feedback gain vector $f \in \mathcal{C}^n$ such that the set of eigenvalues of the closed-loop matrix $A - bf^T$ is $\mathcal{P}$.*

(Note that we use f^T, although f may be complex, since this simplifies the formulas.)

It is well known that for an arbitrary pole set $\mathcal{P}$, the feedback f always exists and is unique if and only if (A, b) is controllable, see, e.g., ([20, page 48, Theorem 2.1]). This problem has been studied extensively. In the literature there are some explicit formulas known for f and the Jordan canonical form of $A - bf^T$ as well as many perturbation results; see [1, 11, 12, 2, 17]. Also, many numerical algorithms have been proposed for this problem; see [2, 4, 10, 14, 15, 16, 18, 8]. In order to analyze the validity of the computed results and the robustness of the solution, it is very important to study the sensitivity of the solution with respect to perturbations in the data. This question has lead to some confusion in the literature [10, 11, 8, 12]. This confusion arises mainly from the fact that there are essentially two different types of results for the SIPP, namely, the feedback gain vector f and the eigenstructure of the closed loop matrix $A - bf^T$. It is possible and actually quite common that, even though f is insensitive to perturbations in the data, the spectrum of $A - bf^T$ is very

*Received by the editors April 22, 1996; accepted for publication (in revised form) by P. Van Dooren August 21, 1997; published electronically March 18, 1998.
http://www.siam.org/journals/simax/19-3/30238.html

†Fakultät für Mathematik, TU Chemnitz, D-09107 Chemnitz, Germany (mehrmann@mathematik.tu-chemnitz.de). The research of this author was partially supported by DFG project Me 790/7-2: Singuläre Steuerungsprobleme.

‡Department of Mathematics, Fudan University, Shanghai 200433, People's Republic of China. Current address: Fakultät für Mathematik, TU Chemnitz, D-09107 Chemnitz, Germany (hxu@mathematik.tu-chemnitz.de). The research of this author was supported by the Alexander von Humboldt Foundation and the Chinese National Natural Foundation.

sensitive, and vice versa. Examples 1 and 4 below demonstrate this phenomenon. It follows that there are also two condition numbers to study here, one for the mapping from $(A, b, \mathcal{P})$ to f and one for the mapping $(A, b, \mathcal{P})$ to the eigenstructure of $A - bf^T$. This observation, and the fact that it is often not explicitely stated which solution of the SIPP problem is considered, explains some of the confusion in the literature. From the point of view of applications, in our opinion the more important problem is to guarantee that the poles of the computed closed loop system $A - bf^T$ are close to the desired ones; the accuracy of f is less important. In particular, in applications such as stabilization it would be fatal if the closed loop poles were made unstable by very small perturbations. To see that this may happen very easily and unexpectedly, consider the following example.

Example 1 (see [12]). Consider the SIPP problem with data

$$A = \operatorname{diag}(1, 2, \ldots, 15), \quad b = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathcal{P} = \{-1, \ldots, -15\}.$$

In this case, both f and $A - bf^T$ can be computed analytically and hence no rounding errors occur in these quantities; see [12] for details. But the eigenvalues of the closed loop systems are so sensitive to perturbations that some of the computed eigenvalues of $A - bf^T$ are in the right half plane.

Unfortunately, this example is not an exception, as was pointed out in [8] and partially proved in [12]; the SIPP problem is usually ill-conditioned. This means that in most cases, in particular if the system size is large ($n > 10$), small perturbations in the data $A, b, \mathcal{P}$ will cause large perturbations in the eigenstructure of $A - bf^T$. In practice it can be expected that if such a feedback is implemented, then the real behavior of the closed loop system is very different than the expected behavior.

Based on these results we have to reconsider the pole placement problem. Since in applications one almost never needs the poles of the closed-loop system in fixed positions but rather in specific regions in the complex plane, it is natural to ask the question of whether we can optimize the conditioning of the problem by varying the choice of poles in the prescribed regions. If we reconsider pole placement in this way, then we have the following problem.

Optimal single-input pole placement (OSIPP). *Given $A \in \mathcal{C}^{n\times n}$, $b \in \mathcal{C}^n$ and a set $\mathcal{D} \subset \mathcal{C}$, find poles $\lambda_1, \ldots, \lambda_n \in \mathcal{D}$, i.e., $\mathcal{P} \subset \mathcal{D}$, such that the SIPP problem is optimally conditioned among all possible choices of $\mathcal{P} \subset \mathcal{D}$.*

To study this problem we first have to discuss what the condition number of the SIPP problem is and how it can be computed or estimated. In particular it would be important for an optimization to have an easily computable quantity or estimate.

In general the condition number of a problem measures the sensitivity of the solution with repect to perturbations in the input data. As we have mentioned above, several quantities can be viewed as solutions of the pole placement problem and different formulas and bounds have been given in recent years; see [2, 8, 11, 17, 12].

We will base our optimization on modifications of the formula for the condition number given in [17]. This formula is very difficult to compute so that an optimization for this condition number seems hopeless. In [12], therefore, based on explicit solution formulas for f and the closed loop eigenvector matrix, slightly different bounds were obtained. We will modify these bounds again to obtain quantities which we can optimize.

To do this we introduce the following notation. The *scaled spectral condition number* of a diagonalizable matrix A is defined as

$$\kappa := \|G\| \, \|G^{-1}\|,$$

where G is the eigenvector matrix of A normalized such that all columns have unit norm. The scaled spectral condition number is equivalent to the optimal spectral condition number; see [5]. Here we use $\|\cdot\|$ to denote an arbitrary consistent norm and use $\|\cdot\|_2$, $\|\cdot\|_F$ to denote the Euclidean and Frobenius norm, respectively. We denote by κ, κ_2, and κ_F the associated scaled spectral condition numbers of $A-bf^T$. By $\Lambda(A)$ we denote the set of eigenvalues of a square matrix A, and by $\sigma_1(B) \geq \cdots \geq \sigma_p(B) \geq 0$, $p = \min\{m, n\}$ we denote the singular values of an $m \times n$ matrix B; see [7]. By $\mathcal{C}_0^+$, $\mathcal{C}^-$, and $\mathcal{C}_{-\rho}^-$ we denote closed right half plane, open left half plane, and the set of complex numbers with real parts not larger than $-\rho$ for $\rho > 0$, respectively. Finally we set e to be the vector of all ones and e_i to be the ith unit vector.

The following perturbation theorem is a combination of two perturbation results given in [12], with slightly modified assumptions.

THEOREM 1.1. *Consider the SIPP problem with data $A, b, \mathcal{P} = \{\lambda_1, \ldots, \lambda_n\}$. Assume that (A, b) is controllable and that the n poles λ_i are distinct. Let $\lambda = [\lambda_1, \ldots, \lambda_n]^T$. Consider also the perturbed problem with data $\hat{A} := A + \delta A$, $\hat{b} := b + \delta b$, and $\delta\lambda = [\delta\lambda_1, \ldots, \delta\lambda_n]^T$. Assume that $(\hat{A}, \hat{b})$ is also controllable and that also the perturbed poles are distinct. Set $\epsilon := \max\{\|[\delta A \ \delta b]\|, \max_i |\delta\lambda_i|\}$ and suppose that*

$$2\epsilon < \min_i \sigma_n \begin{bmatrix} A - \lambda_i I & b \end{bmatrix} =: \sigma_\lambda.$$

Let f, $\hat{f} := f + \delta f$ be the feedback gains of the original and perturbed problems, respectively, then

$$\|\delta f\|_2 \leq c_f := \frac{2\sqrt{2n}}{\sigma_\lambda} \epsilon \hat{\kappa}_2 \sqrt{1 + \|f\|_2^2} \max_i \sqrt{\left(\frac{\|\hat{A} - \hat{\lambda}_i I\|_2}{\|\hat{b}\|_2} \right)^2 + 1}, \tag{2}$$

where $\hat{\kappa}_2$ is the scaled spectral condition number of $\hat{A} - \hat{b}\hat{f}^T$.

Furthermore, for each eigenvalue μ_i of the computed closed-loop matrix $A - b\hat{f}^T$, there exists a corresponding eigenvalue λ_i of the desired (unperturbed) closed-loop system such that

$$|\mu_i - \lambda_i| \leq c_e := \left(1 + \hat{\kappa}_2 \sqrt{1 + \|\hat{f}\|_2^2} \right) \epsilon. \tag{3}$$

Proof. The proof is easily obtained from the proofs of Theorems 7 and 8 in [12], using the slightly different assumptions on δA, δb, $\delta\lambda$. The estimates (2), (3) are obtained analogously. ∎

We see from Theorem 1.1 that several factors contribute to the perturbation bounds and thus can be considered to create large perturbations in f and/or the closed-loop eigenvalues. The main factors in the bound (2) are the quantity σ_λ and the term $\hat{\kappa}\sqrt{1 + \|f\|_2^2}$. In the following we restrict our optimization to the second factor. This may lead to an overestimation of the bound in the optimal case but it simplifies the optimization and can be justified as follows. The term σ_λ reflects the distance to uncontrollability $d_{uc}(A, b) = \min_{s \in \mathcal{C}} \sigma_n[A - sI, b]$, which is independent

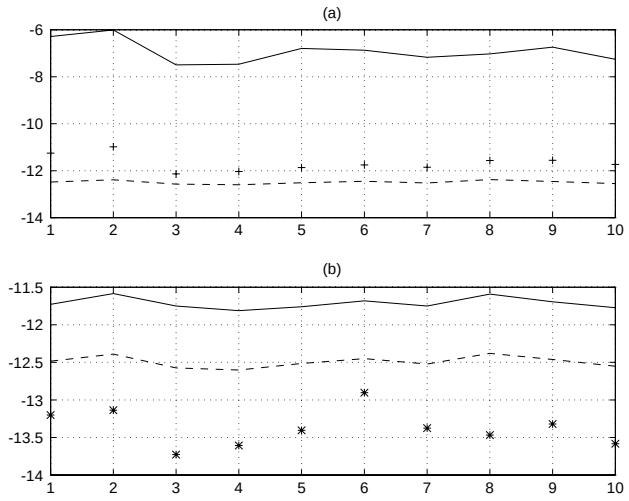


FIG. 1. *Error and bounds in Example* 2: (a) *feedback gain:* $\log(\|f - \hat{f}\|) : +++$, $\log(c_f) : —$, $\log(eps \cdot \mathcal{S}) : - - -$; (b) *closed-loop poles*: $\log(eig_e) : ***$, $\log(c_e) : —$, $\log(eps \cdot \mathcal{S}) : - - -$.

of the choice of poles. So, we can replace σ_λ in (2) to obtain an upper bound. But it should be noted that $d_{uc}(A, b)$ can be much smaller than σ_λ, see, e.g., [3]. If $d_{uc}(A, b)$ is reasonably large, then replacing σ_λ by $d_{uc}(A, b)$ will have only a small effect on the bound. If, however, $d_{uc}(A, B)$ is very small, then this may lead to an overestimation of the condition number. This effect is demonstrated in some of our numerical examples below. But in this case we know that the problem is very close to a problem which is not controllable. This is a critical situation in practice and it may be reasonable to modify the model in such a case. Consider now the term $\hat{\kappa}\sqrt{1 + \|f\|_2^2}$, which governs the perturbations in the computed closed-loop eigenvalues. This term (if large) also makes the bound (2) very large. If this term can be made small by the choice of poles and if $d_{uc}(A, b)$ is reasonably large, then both bounds (2) and (3) are small and we can expect that f can be computed accurately and that the closed loop eigenvalues are robust. We will therefore optimize the quantity

$$\mathcal{S} := \kappa\sqrt{1 + \|f\|^2} \tag{4}$$

to improve the conditioning via the choice of poles. If necessary, we use the notation $\mathcal{S}_2 := \kappa_2\sqrt{1 + \|f\|_2^2}$ and $\mathcal{S}_F := \kappa_F\sqrt{1 + \|f\|_2^2}$. In order to illustrate that $\mathcal{S}$ catches the qualitative behavior of the errors well, we will consider the following numerical tests.

All computations in this paper were carried out in Matlab Version 4.2 on a pentium-s PC with machine precision $eps = 2.22 \times 10^{-16}$. Random matrices are created with the Matlab *rand* function and uniform distribution. In the figures below we depict c_f as in (2), c_e as in (3) with $\epsilon = \max\{\|[A\ b]\|, \max_i |\lambda_i|\}eps$, and by eig_e we denote the maximal error between an eigenvalue of the computed closed-loop matrix and the associated pole in $\mathcal{P}$.

Example 2. In this example we constructed ten problems with $A \in \mathcal{R}^{30\times 30}$, $b \in \mathcal{R}^{30}$ with elements chosen randomly in $[-1, 1]$, and for each of these problems we chose 50 different (exact) feedback gains $f \in \mathcal{R}^{30}$ with random elements in $[-10, 10]$. For each of these 500 problems the chosen poles are the computed eigenvalues of $A - bf^T$ (via the Matlab *eig* function). Then we computed the feedback gain $\hat{f}$ with

TABLE 1
min *and* max *of* $\log(c_f/\|f-\hat{f}\|)$.

Problem	1	2	3	4	5	6	7	8	9	10
min	3.6	3.3	3.6	3.3	3.6	3.8	3.4	3.4	3.6	3.1
max	6.5	7.0	5.9	6.1	7.1	6.8	6.4	5.9	6.0	5.7

TABLE 2
min *and* max *of* $\log(c_e/eig_e)$.

Problem	1	2	3	4	5	6	7	8	9	10
min	-0.1	-0.1	1.6	1.2	1.3	0.9	1.0	1.1	0.9	0.8
max	3.2	3.1	3.2	2.9	2.1	3.1	2.8	3.3	2.8	2.6

these poles by using Miminis and Paige's *sevas* Matlab code (cf. [14]). In Figure 1 for each of the 10 pairs (A, b) we display the arithmetic means of logarithms of c_e, eig_e, c_f and $\|f-\hat{f}\|$ taken over the 50 experiments and compare it with $eps \cdot \mathcal{S}$. We also display the minimum and maximum distances between the orders of c_f and $\|f-\hat{f}\|$, e_f, and eig_e in Tables 1 and 2, respectively, among the 50 experiments for each pair (A, b).

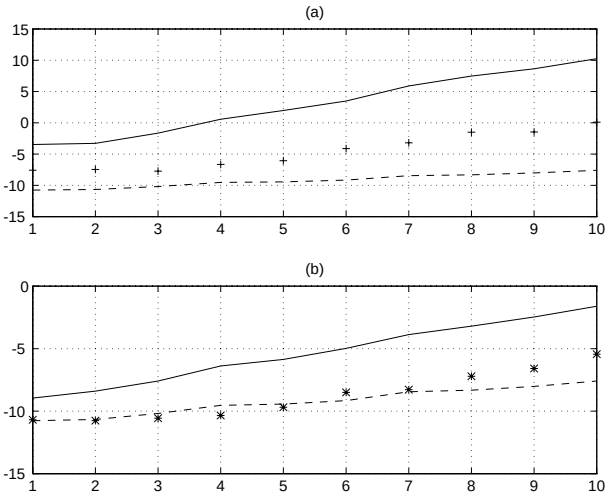


FIG. 2. *Error and bounds in Example* 3: (a) *feedback gain:* $\log(\|f-\hat{f}\|) : + + +$, $\log(c_f) :$ —, $\log(eps \cdot \mathcal{S}) :$ - - -; (b) *closed-loop poles:* $\log(eig_e) : * * *$, $\log(c_e) :$ —, $\log(eps \cdot \mathcal{S}) :$ - - -.

We see that the bounds and also the term $\mathcal{S}$ describe the qualitative behavior of the errors quite well, although the bounds (2) and (3) sometimes tend to be too pessimistic. In general we cannot expect to see more than the qualitative behavior, since we have omitted terms in the bounds. In Example 2 we have that σ_λ contributes a factor 10^2, which almost explains the difference of the bounds.

The estimate of the errors in the closed-loop eigenvalues is in general better than that of the feedback gain, since the factors do not occur.

The reason for the negative numbers in Table 2 is an underestimation of ϵ.

Example 3. The conditioning becomes worse if we modify Example 2 by diagonal scaling of A. Let $A := D\tilde{A}D^{-1}$, $D = \mathrm{diag}(1, 2^{\frac{m+2}{3}}, \ldots, 30^{\frac{m+2}{3}})$ with $\tilde{A}$ as in Example 2. Here we let m take the values 1 to 10, and the poles are formed as before. The

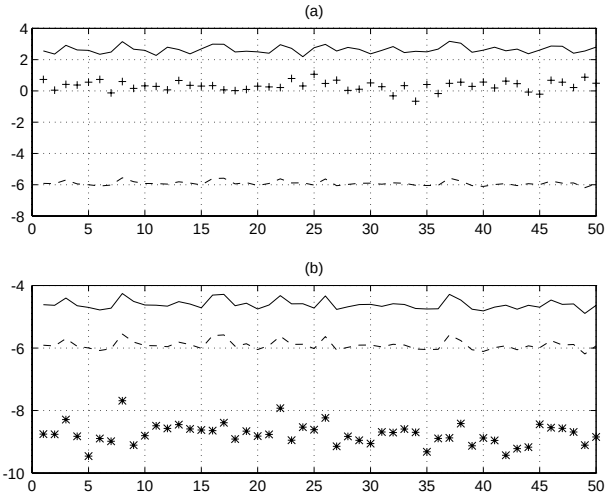


FIG. 3. *Error and bounds in Example* 4: (a) *feedback gain:* $\log(\|f - \hat{f}\|) : + + +$, $\log(c_f) :$ —, $\log(eps \cdot \mathcal{S}) :$ - - -; (b) *closed-loop poles:* $\log(eig_e) : * * *$, $\log(c_e) :$ —, $\log(eps \cdot \mathcal{S}) :$ - - -.

errors and bounds for this example are in given Figure 2.

Our next example demonstrates that it is not sufficient to consider the accuracy of the feedback gain f alone. In this case the computed closed-loop eigenvalues are much more accurate than the computed f.

Example 4. Let

$$A = Q^T \begin{bmatrix} -1 & -1 & \cdots & -1 \\ 1 & -1 & \ddots & \vdots \\ \mathbf{O} & \ddots & \ddots & \vdots \\ & & 1 & -1 \end{bmatrix} Q \in \mathcal{R}^{30\times 30}, \quad b = Q^T e_1 \in \mathcal{R}^{30},$$

where Q is a random orthogonal matrix, $f = 10^5 Q^T f_1$, and $f_1 \in \mathcal{R}^{30}$ has elements randomly selected in $[-1, 1]$. We chose 50 random vectors f_1 and produced the poles as in Example 2.

Figure 3 shows that, even though on the average the computed f has no correct digit, the closed loop eigenvalues still carry essentially eight correct digits.

2. Minimization of $\mathcal{S}$. In view of the discussion in the previous section, we now consider the minimization of $\mathcal{S}$, as defined in (4), when the closed-loop eigenvalues are allowed to vary in a given set $\mathcal{D} \subset \mathcal{C}$.

In the following we restrict our minimization to the case that all elements in $\mathcal{P}$ are distinct so that $A - bf^T$ is diagonalizable. Although condition numbers are also interesting in the degenerate case, they are more complicated and usually the conditioning is much worse ([19, pages 87–90]).

If we consider the SIPP problem with data A, b, λ, where (A, b) is controllable and the components of λ are distinct, then explicit formulas for the solution of the SIPP problem are known; see [1, 4, 12]. For the case of distinct poles we have the following formula from [12]. Let

$$G = [u_1, \ldots, u_n], \quad \|u_i\|_2 = 1, \quad i = 1, \ldots, n, \tag{5}$$

and $\alpha = [\alpha_1, \ldots, \alpha_n]$ such that for each i, $[u_i^T, -\alpha_i]^T$ is nonzero and satisfies

$$[A - \lambda_i I, b] \begin{bmatrix} u_i \\ -\alpha_i \end{bmatrix} = 0. \tag{6}$$

Then G is nonsingular,

$$f^T = \alpha^T G^{-1} = e_n^T [b, Ab, \ldots, A^{n-1}b]^{-1} \prod_{i=1}^{n} (A - \lambda_i I), \tag{7}$$

$$A - bf^T = G \operatorname{diag}(\lambda_1, \ldots, \lambda_n) G^{-1},$$

and

$$\kappa = \|G\| \, \|G^{-1}\|. \tag{8}$$

In principle we could employ nonlinear optimization methods to compute the required minimum, but considering the explicit formulas (5), (6), and (7), this is a very difficult problem, in particular when n is large.

We will now discuss some cases where we can give explicit formulas for $\|f\|$ and κ and thus, also, the optimization problem becomes much simpler.

THEOREM 2.1. *Let $A = \Gamma := \operatorname{diag}(\gamma_1, \ldots, \gamma_n)$, $b = e$, and (Γ, e) be controllable, let $\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\}$ be a pole set with distinct elements and $\Lambda(A) \cap \mathcal{P} = \emptyset$. Then*

$$\|f\|_2^2 = \sum_{i=1}^{n} \frac{\prod_{k=1}^{n} |\gamma_i - \lambda_k|^2}{\prod_{k=1, k\neq i}^{n} |\gamma_i - \gamma_k|^2}, \tag{9}$$

$$\kappa_F^2 = n \sum_{i,j=1}^{n} \frac{\sum_{l=1}^{n} \prod_{k=1, k\neq l}^{n} |\lambda_i - \gamma_k|^2}{\prod_{k=1, k\neq i}^{n} |\lambda_i - \lambda_k|^2} \frac{\prod_{k=1, k\neq i}^{n} |\gamma_j - \lambda_k|^2}{\prod_{k=1, k\neq j}^{n} |\gamma_j - \gamma_k|^2}. \tag{10}$$

Proof. Define the Cauchy matrix $C = [c_{ij}]_{n\times n}$ with $c_{ij} = \frac{1}{\gamma_i - \lambda_j}$. Then it follows from the inversion formula for Cauchy matrices in [6] that $C^{-1} = -W\, C^T\, H$, where $W := \operatorname{diag}(w_1, , \ldots, w_n)$, $H := \operatorname{diag}(h_1, \ldots, h_n)$, and

$$w_i := \frac{\prod_{k=1}^{n} (\lambda_i - \gamma_k)}{\prod_{k=1, k\neq i}^{n} (\lambda_i - \lambda_k)}, h_i := \frac{\prod_{k=1}^{n} (\gamma_i - \lambda_k)}{\prod_{k=1, k\neq i}^{n} (\gamma_i - \gamma_k)}, \qquad i = 1, \ldots, n. \tag{11}$$

Applying formulas (5)–(8) to these special data A, b and $\mathcal{P}$ we get $f^T = e^T C^{-1}$. Using the formula for C^{-1} we get $f_i = h_i$, $k = 1, \ldots, n$, where f_i is the ith component of f. With the formulas of h_i in (11) we obtain (9).

Using the definition of κ in (8) and noticing that C is an eigenvector matrix of $A - bf^T = \Gamma - ef^T$, we only need to normalize the columns of C to one. Let $U := \operatorname{diag}(\mu_1, \ldots, \mu_n)$, where

$$\mu_i := \sqrt{\sum_{k=1}^{n} \frac{1}{|\gamma_k - \lambda_i|^2}} = \sqrt{\frac{\sum_{l=1}^{n} \prod_{k=1, k\neq l}^{n} |\gamma_k - \lambda_i|^2}{\prod_{k=1}^{n} |\gamma_k - \lambda_i|^2}},$$

and let $C_0 := CU^{-1}$; then we get $\kappa_F^2 = \|C_0\|_F^2 \, \|{C_0}^{-1}\|_F^2$. Clearly $\|C_0\|_F^2 = n$, and using the formulas for C^{-1} and U we get (10). ∎

Remark 1. Let $\psi(t) = \prod_{k=1}^{n}(t-\lambda_k)$ be the characteristic polynomial of $\Gamma - ef^T$. It is easy to check with $f_i = h_i$ as in (11) that the components of f are just the coefficients of the Lagrange interpolating polynomial for the n points $\{\gamma_k, \psi(\gamma_k)\}_{k=1}^{n}$, i.e., the polynomial $\eta(t) = \sum_{k=1}^{n} f_k \prod_{l=1,l\neq k}^{n}(t-\gamma_l)$ satisfies $\eta(\gamma_k) = \psi(\lambda_k)$. Moreover we have $\psi(t)-\eta(t) = \prod_{k=1}^{n}(t-\gamma_k) =: \phi(t)$, which is just the characteristic polynomial of Γ.

In Theorem 2.1 we have obtained κ_F and $\|f\|_2$ in terms of the pole set and the spectrum of Γ. The evaluation of the polynomial $\|f\|_2^2$ and the rational function κ_F^2 in an optimization code is relatively simple; however, there are still difficulties when n, the size of the problem, is large. The second difficulty in employing an optimization procedure is the selection of the initial value. A bad initial value will lead to an extremely large $\mathcal{S}$ and for large n this may lead to overflow in the computations. So even if $\mathcal{P}$ exists such that the given SIPP problem is well conditioned, it will be difficult to start the optimization procedure. The third difficulty is that for some systems (A, b) and sets $\mathcal{D}$, even the OSIPP problems is ill conditioned, as we will show in Example 5. In such a case there is no need to use an optimization procedure. Theorem 2.1 also shows that the minimum of $\mathcal{S}$ approaches infinity if some $|\lambda_i| \to \infty$, since in this case $\|f\| \to \infty$ and $\kappa \geq 1$, so $\mathcal{S} \to \infty$.

The following result shows that if A has all distinct and simple eigenvalues, then it is usually sufficient to restrict our discussion to the special case $(\Gamma, e, \mathcal{P})$.

THEOREM 2.2. *Let (A, b) be controllable, $A = X\Gamma X^{-1}$, $\Gamma := \operatorname{diag}(\gamma_1, \ldots, \gamma_n)$, and let $\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\}$ have distinct elements. Denote by f_d, f the unique feedback gains of $(A, b, \mathcal{P})$ and $(\Gamma, e, \mathcal{P})$, respectively, and by κ_F^d, κ_F denote the associated scaled spectral condition numbers of $A - bf_d^T$ and $\Gamma - ef^T$ in Frobenius norm. Let $\tilde{b} = \begin{bmatrix} \tilde{b}_1 & \cdots & \tilde{b}_n \end{bmatrix}^T := X^{-1}b$ and $B := \operatorname{diag}(\tilde{b}_1, \ldots, \tilde{b}_n)$. Then*

$$\frac{\|f\|_2}{\|XB\|_2} \leq \|f_d\|_2 \leq \|(XB)^{-1}\|_2 \|f\|_2 \tag{12}$$

and

$$\frac{\kappa_F}{\sqrt{n}\,\|XB\|_2\,\|(XB)^{-1}\|_2} \leq \kappa_F^d \leq \sqrt{n}\,\|XB\|_2\,\|(XB)^{-1}\|_2 \kappa_F. \tag{13}$$

Proof. Let

$$A - bf_d^T = G\Lambda G^{-1}, \quad \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n), \quad \kappa_F^d = \|G\|_F \|G^{-1}\|_F,$$

i.e., the columns of G have unit norm. With $A = X\Gamma X^{-1}$ and B defined as above, we have

$$\Gamma - Bef_d^T X = X^{-1}G\Lambda G^{-1}X.$$

Since (A, b) is controllable if and only if (Γ, e) is, we obtain (see also [12]) $\tilde{b}_i \neq 0$, $i = 1, \ldots, n$, so B is nonsingular. Performing a similarity transformation with B^{-1}, B, and using that B and Γ are diagonal, we have

$$\Gamma - ef_d^T XB = (XB)^{-1}G\Lambda((XB)^{-1}G)^{-1}.$$

By the uniqueness of the feedback gain, we then have $f^T = f_d^T XB$, which implies (12). Let C_0 be defined as in the proof of Theorem 2.1; then $C_0 = (XB)^{-1}GZ$, for

some diagonal matrix $Z = \operatorname{diag}(z_1, \ldots, z_n)$ so that the columns of C_0 have unit norm. So we have

$$\sqrt{n} = \|C_0\|_F \geq \frac{\|GZ\|_F}{\|XB\|_2} = \frac{\sqrt{\sum_{k=1}^n |z_k|^2}}{\|XB\|_2} \geq \frac{\max_k |z_k|}{\|XB\|_2} = \frac{\|Z\|_2}{\|XB\|_2}, \tag{14}$$

i.e., $\|Z\|_2 \leq \sqrt{n}\,\|XB\|_2$. Similarly, by using $G = XBC_0Z^{-1}$, we get $\|Z^{-1}\|_2 \leq \sqrt{n}\|(XB)^{-1}\|_2$. Since $G^{-1} = ZC_0{}^{-1}(XB)^{-1}$, we obtain from (14) that

$$\frac{\|C_0^{-1}\|_F}{\sqrt{n}\|(XB)^{-1}\|_2\,\|XB\|_2} \leq \|G^{-1}\|_F \leq \sqrt{n}\,\|XB\|_2\,\|(XB)^{-1}\|_2\|C_0^{-1}\|_F.$$

Using $\kappa_F^d = \sqrt{n}\|G^{-1}\|_F$, $\kappa_F = \sqrt{n}\|C_0{}^{-1}\|_F$, we obtain (13). ∎

We see from this result that since $\|XB\|\,\|(XB)^{-1}\|$ is independent of the chosen poles, we can restrict ourselves to the special case $(\Gamma, e, \mathcal{P})$. It is obvious that (12) and (13) and also all subsequent bounds can be extended to the case when (A, b) is controllable and A is diagonalizable.

To analyze the problem further in this case, we will give several bounds for $\mathcal{S}$, as defined in (4).

THEOREM 2.3. *Let $\Gamma := \operatorname{diag}(\gamma_1, \ldots, \gamma_n)$, let (Γ, e) be controllable, and let $\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\}$ with distinct elements. Suppose that $\lambda(\Gamma) \cap \mathcal{P} = \emptyset$ and set $d_u = \max_{i,j} |\gamma_i - \lambda_j|$, $d_l = \min_{i,j} |\gamma_i - \lambda_j|$. Furthermore, set $w := [w_1, \ldots, w_n]^T$ with $w_i = \frac{\prod_{k=1}^n (\lambda_i - \gamma_k)}{\prod_{k=1, k\neq i}^n (\lambda_i - \lambda_k)}$. Then*

$$n\,\frac{\|w\|_2\,\|f\|_2\,\sqrt{1 + \|f\|_2^2}}{d_u^2} \leq \mathcal{S}_F \leq n\,\frac{\|w\|_2\,\|f\|_2\,\sqrt{1 + \|f\|_2^2}}{d_l^2}. \tag{15}$$

Proof. Considering the formulas for κ_F and f in Theorem 2.1 and (11), we obtain

$$\begin{aligned}
\kappa_F^2 &= n \sum_{i,j=1}^n \frac{\sum_{l=1}^n \prod_{k=1,k\neq l}^n |\lambda_i - \gamma_k|^2}{\prod_{k=1,k\neq i}^n |\lambda_i - \lambda_k|^2} \frac{\prod_{k=1,k\neq i}^n |\gamma_j - \lambda_k|^2}{\prod_{k=1,k\neq j}^n |\gamma_j - \gamma_k|^2} \\
&= n \sum_{i,j=1}^n \frac{\prod_{k=1}^n |\lambda_i - \gamma_k|^2}{\prod_{k=1,k\neq i}^n |\lambda_i - \lambda_k|^2} \left(\sum_{l=1}^n \frac{1}{|\lambda_i - \gamma_l|^2} \right) \frac{1}{|\gamma_j - \lambda_i|^2} \frac{\prod_{k=1}^n |\gamma_j - \lambda_k|^2}{\prod_{k=1,k\neq j}^n |\gamma_j - \gamma_k|^2} \\
&= n \sum_{i,j=1}^n \left(\sum_{l=1}^n \frac{1}{|\lambda_i - \gamma_l|^2} \right) \frac{1}{|\gamma_j - \lambda_i|^2} |w_i|^2 |f_j|^2.
\end{aligned}$$

Since for all i, j we have $d_l \leq |\lambda_i - \gamma_j| \leq d_u$, it follows that

$$\frac{n}{d_u^4} \leq \left(\sum_{l=1}^n \frac{1}{|\lambda_i - \gamma_l|^2} \right) \frac{1}{|\gamma_j - \lambda_i|^2} \leq \frac{n}{d_l^4},$$

and hence

$$\frac{n^2}{d_u^4} \sum_{i,j=1}^n |w_i|^2 |f_j|^2 \leq \kappa_F^2 \leq \frac{n^2}{d_l^4} \sum_{i,j=1}^n |w_i|^2 |f_j|^2.$$

Thus,

$$\frac{n}{d_u^2}\,\|w\|_2\,\|f\|_2 \le \kappa_F \le \frac{n}{d_l^2}\,\|w\|_2\,\|f\|_2\,,$$

and multiplying with $\sqrt{1+\|f\|_2^2}$ yields the conclusion. ∎

The quantities d_l, d_u are the smallest and largest distances between the sets $\Lambda(\Gamma)$ and $\mathcal{P}$. If $d_l << d_u$, in particular when d_l is very small, the upper bound in (15) will usually be an overestimate. But if d_u/d_l is not too large, (15) will be a good estimate for $\mathcal{S}$.

Note that w is the feedback gain for a SIPP problem with $A = \operatorname{diag}(\lambda_1,\dots,\lambda_n)$, $b = e$, and the pole set $\mathcal{P} = \{\gamma_1,\dots,\gamma_n\}$. So it has a similar interpretation as f in Remark 1 but in general there is no explicit relationship between $\|w\|$ and $\|f\|$. However, if $\mathcal{P}$ is selected in a particular way, we can get $\|w\| = \|f\|$.

COROLLARY 2.4. *Let* $\Gamma := \operatorname{diag}(\gamma_1,\dots,\gamma_n)$ *and let* (Γ, e) *be controllable. If* $\mathcal{P} = \{-\overline{\gamma}_1,\dots,-\overline{\gamma}_n\}$, $\{\overline{\gamma}_1,\dots,\overline{\gamma}_n\}$, *or* $\{-\gamma_1,\dots,-\gamma_n\}$, *then* $\|w\|_2 = \|f\|_2$ *and*

$$\frac{n}{d_u^2}\,\|f\|_2^2\sqrt{1+\|f\|_2^2} \le \mathcal{S}_F \le \frac{n}{d_l^2}\,\|f\|_2^2\sqrt{1+\|f\|_2^2}. \tag{16}$$

Proof. We consider just the case when $\mathcal{P} = \{-\overline{\gamma}_1,\dots,\overline{\gamma}_n\}$; the other two cases are analogous.

Now $\lambda_i = -\overline{\gamma}_i$, $i = 1,\dots,n$, so

$$\begin{aligned} w_i &= \frac{\prod_{k=1}^n(\lambda_i-\gamma_k)}{\prod_{k=1,k\neq i}^n(\lambda_i-\lambda_k)} = \frac{\prod_{k=1}^n(-\overline{\gamma}_i-\gamma_k)}{\prod_{k=1,k\neq i}^n(-\overline{\gamma}_i+\overline{\gamma}_k)} \\ &= -\overline{\left(\frac{\prod_{k=1}^n(\gamma_i+\overline{\gamma}_k)}{\prod_{k=1,k\neq i}^n(\gamma_i-\gamma_k)}\right)} = -\overline{\left(\frac{\prod_{k=1}^n(\gamma_i-\lambda_k)}{\prod_{k=1,k\neq i}^n(\gamma_i-\gamma_k)}\right)} = -\overline{f}_i. \end{aligned}$$

Hence $\|w\|_2 = \|f\|_2$, and then (16) follows from (15). ∎

Since the solution of the SIPP problem consists of the computation of f, it is natural that we try to estimate κ in terms of $\|f\|$.

Such a result is useless for an optimization procedure since the poles are fixed, but it is valuable in some typical cases arising in applications, i.e., the cases when $\mathcal{P}$ is chosen such that the eigenvalues of A are reflected at the real axis, imaginary axis, or origin, as for example in the well-known Lyapunov method; see, e.g., [9]. A (lower) bound for $\mathcal{S}$ in terms of $\|f\|$ is the following.

THEOREM 2.5. *Consider the SIPP problem with data A, b,* $\mathcal{P}$, *where* (A, b) *is controllable and the poles in* $\mathcal{P}$ *are distinct. Then*

$$\mathcal{S}_2 \ge \frac{\|b\|_2}{\sqrt{\sum_{i=1}^n \|A-\lambda_i I\|_2^2}}\,\|f\|_2\sqrt{1+\|f\|_2^2}. \tag{17}$$

Proof. In this case we can apply formulas (5)–(8). From $f^T = \alpha^T G^{-1}$ we get $\|f\|_2 \le \|\alpha\|_2 \left\|G^{-1}\right\|_2$, i.e. ,

$$\|G^{-1}\|_2 \ge \frac{\|f\|_2}{\|\alpha\|_2}.$$

Now α_i, the ith component of α together with u_i, the ith column of G, with $\|u_i\|_2 = 1$, satisfies

$$[A-\lambda_i I, b]\begin{bmatrix} u_i \\ -\alpha_i \end{bmatrix} = 0, \quad \forall i = 1,\dots,n.$$

Thus, it follows that

$$\|b\|_2^2 \, |\alpha_i|^2 = \|(A - \lambda_i I)u_i\|_2^2 \le \|A - \lambda_i I\|_2^2 \, \|u_i\|_2^2 = \|A - \lambda_i I\|_2^2 ,$$

which means that $|\alpha_i| \le \frac{\|A-\lambda_i I\|_2}{\|b\|_2}$. So

$$\|\alpha\|_2 \ \le \frac{\sqrt{\sum_{i=1}^n \|A - \lambda_i I\|_2^2}}{\|b\|_2},$$

and by using $\|G\|_2 \ge \max_i \|u_i\|_2 = 1$, we finally get

$$\kappa_2 \ge \|G\|_2 \, \|G^{-1}\|_2 \ge \|G\|_2 \, \frac{\|f\|_2}{\|\alpha\|_2} \ge \frac{\|b\|_2 \, \|f\|_2}{\sqrt{\sum_{i=1}^n \|A - \lambda_i I\|_2^2}}. \qquad \square$$

This lower bound may be very weak. For example, if all poles are selected in a small neighborhood of a single point λ_i, then $\|f\|_2$ is bounded but κ will be very large. But nevertheless, Theorem 2.5 gives a cheap way to estimate $\mathcal{S} = \kappa\sqrt{1 + \|f\|^2}$. If, for example, the computed $\|f\|$ is large, say $\|f\| > 1/\sqrt{eps}$, where *eps* is the machine epsilon, then $\mathcal{S} > c/eps$ for some constant c. In this case, we can expect that the computed results have lost all significant digits. Consider again special cases where we have explicit solutions.

THEOREM 2.6. *Let* $\Gamma := \operatorname{diag}(\gamma_1, \ldots, \gamma_n)$*, with* $\gamma_j \in \mathcal{C}_0^+$*,* $j = 1, \ldots, n$ *and assume that* (Γ, e) *is controllable.*

(a) *If we require that* $\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\} \subset \mathcal{C}_{-\rho}^-$*, i.e.,* $\operatorname{Re}\lambda_j \le -\rho$*,* $j = 1, \ldots, n$ *for a given real number* $\rho > 0$*, then* $\min \|f\|_2$ *is obtained when* $\operatorname{Re}\lambda_j = -\rho$*, for all* $j = 1, \ldots, n$*.*

(b) *If the* γ_j *are such that* $\operatorname{Re}\gamma_j + \rho \ge |\operatorname{Im}\gamma_j|$*, for* $j = 1, \ldots, n$*, and we require that the set of poles is as in* (a) *and closed under conjugation, then*

$$\min \|f\|_2 = \sqrt{\sum_{j=1}^n \frac{|\gamma_j + \rho|^{2n}}{\prod_{k=1, k\ne j}^n |\gamma_j - \gamma_k|^2}}, \tag{18}$$

and the corresponding optimal poles satisfy $\lambda_j = -\rho$*, for all* $j = 1, \ldots, n$*.*

Proof. (a) Let $\gamma_j = a_j + ib_j$, $\lambda_j = x_j + iy_j$, where a_j, b_j, x_j, y_j are real. Then

$$\|f\|_2^2 = \sum_{j=1}^n \frac{\prod_{k=1}^n |\gamma_j - \lambda_k|^2}{\prod_{k=1, k\ne j}^n |\gamma_j - \gamma_k|^2} = \sum_{j=1}^n \frac{\prod_{k=1}^n ((a_j - x_k)^2 + (b_j - y_k)^2)}{\prod_{k=1, k\ne j}^n ((a_j - a_k)^2 + (b_j - b_k)^2)}.$$

Since $a_j \ge 0$, $x_j \le -\rho$ for all j, a necessary condition for a minimum is that $(a_j - x_k)^2$ is minimal for all k, which is clearly the case if $x_k = -\rho$, for all $k = 1, \ldots, n$.

(b) From (a) we obtain that at a minimum all poles have real part $-\rho$. Suppose that, at the minimum, there exists a pole with nonzero imaginary part $\lambda_s = -\rho + iy_s$. Since the pole set is closed under conjugation, we obtain, for each γ_j,

$$\begin{aligned} |\gamma_j - \lambda_s|^2 |\gamma_j - \bar\lambda_s|^2 &= ((a_j + \rho)^2 + (b_j - y_s)^2)((a_j + \rho)^2 + (b_j + y_s)^2) \\ &= y_s^4 + 2((a_j + \rho)^2 - b_j^2) y_s^2 + ((a_j + \rho)^2 + b_j^2)^2. \end{aligned}$$

By assumption, $a_j + \rho \ge |b_j|$, and thus we have

$$\min_{y_s} |\gamma_j - \lambda_s|^2 |\gamma_j - \bar\lambda_s|^2 = ((a_j + \rho)^2 + b_j^2)^2,$$

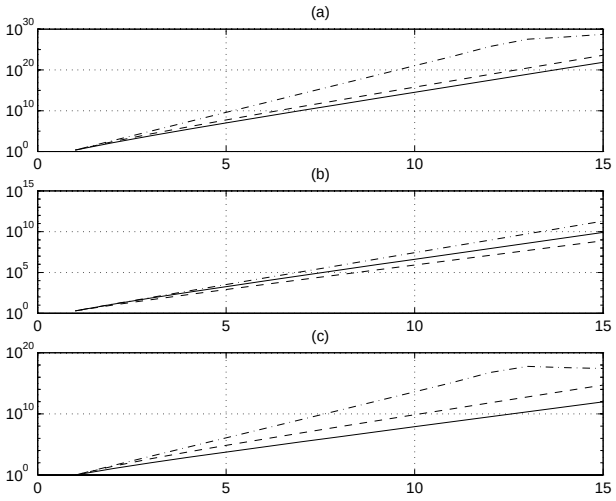


FIG. 4. *Conditioning in Example* 5: (a) $\mathcal{S}$; (b) $\|f\|$; (c) κ; *optimal:* —, $\mathcal{P}_1$: - - -, $\mathcal{P}_2$: – · –·

i.e., the minimum occurs for $y_s = 0$.

Since each component $|f_k|$ must include a factor of the form $|\gamma_k - \lambda_s||\gamma_k - \bar{\lambda}_s|$, to minimize $\|f\|_2$ we must have $y_s = 0$, which is a contradiction to our assumption. Consequently we have $\lambda_j = -\rho$ for all j and hence we obtain (18). □

Part (b) of Theorem 2.6 shows that, in a particular situation, to get a minimal $\|f\|$ is equivalent to getting the worst conditioning for the closed-loop matrix (in the sense of eigenvalue perturbation theory) since we have to place all poles in one point and the controllability forces the closed-loop matrix to be similar to an $n \times n$ Jordan block. This result also explains the extreme ill conditioning of Example 1.

We also see that, in the situation of Theorem 2.6 (a), at least half of the variables can be removed and that the minimization problem for $\|f\|$ is restricted to a line rather than a half plane. In this situation, suppose that $\mathcal{P}_0 \subset \mathcal{C}^-_{-\rho}$ is a pole set that minimizes $\mathcal{S}$ and let f_0 be the feedback gain obtained with $\mathcal{P}_0$. Then with Theorems 2.5 and 2.6 we obtain

$$\min_{\mathcal{P}\subset\mathcal{C}^-_{-\rho}} \mathcal{S} \geq c\,\|f_0\|^2 \geq c \min_{\mathcal{P}\subset\{-\rho+yi|y\in\mathcal{R}\}} \|f\|^2, \tag{19}$$

for a constant c determined by A, b, $\mathcal{P}_0$. Thus, if $\min\|f\|$ is large, then the OSIPP problem is incurably ill conditioned and we cannot hope to improve the problem by choosing the poles. Consider the following example.

Example 5. Let $A = \mathrm{diag}(1, \ldots, n)$, $b = e$, $\mathcal{P} \subset \mathcal{C}^-_{-1}$, and let n vary from 1 to 15. We used a heuristic "random search" algorithm to choose the set $\mathcal{P}_0$ that minimizes $\mathcal{S}$ and determine a minimal value for $\mathcal{S}$. The resulting condition numbers are shown in Figure 4, as well as the related numbers κ and $\|f\|$. For comparison we also display the three numbers $\|f\|$, κ, and $\mathcal{S}$ for the pole sets $\mathcal{P}_1 = \{-1 - \frac{n-1}{2}i, -1 - \frac{n-3}{2}i, \ldots, -1 + \frac{n-3}{2}i, -1 + \frac{n-1}{2}i\}$ and $\mathcal{P}_2 = \{-1, -2, \ldots, -n\}$, respectively.

We see that the magnitude of $\mathcal{S}$ can be reduced, but even so, when $n = 11$, $\mathcal{S}_{opt} = 10^{16}$. For $n = 15$, $\mathcal{S}_{opt} = 7.5 \times 10^{21}$ and 10 eigenvalues of $A - bf^T$ are in $\mathcal{C}^+_0$. So for $n \geq 15$, it is impossible to place the poles to the left of the line $-1 + yi$ via a numerical procedure. We can also check the ill conditioning by the results in Theorems 2.5 and 2.6. Actually, for $n = 15$, $\min\|f\| = 3.45 \times 10^8$.

TABLE 3
Eigenvalue error and eps · $\mathcal{S}$.

n	1	2	3	4	5	6	7
eps · $\mathcal{S}$	5.0e-16	3.3e-14	1.6e-12	6.5e-11	2.3e-9	7.9e-8	2.5e-6
E_n	0	1.8e-15	6.7e-14	1.8e-12	6.7e-11	1.2e-10	3.7e-9
8	9	10	11	12	13	14	15
8.0e-5	0.0024	0.0741	2.2217	66.0347	2.0e+3	5.7e+4	1.7e+6
3.6e-8	1.1e-7	3.4e-5	3.5e-4	4.2e-3	6.0e-2	0.6371	7.132

To illustrate again the importance of $\mathcal{S}$, in Table 3 we list *eps*·$\mathcal{S}$ and the eigenvalue errors for the poles obtained from our heuristic search algorithm for different n. Here $E_n = \max_j |\mu_j - \lambda_j|$, where $\Lambda(A - b\tilde{f}^T) = \{\mu_1, \ldots, \mu_n\}$, and $\tilde{f}$ is the computed feedback with these poles. We see that the error in the eigenvalues grows roughly in the same way as *eps* · $\mathcal{S}$. Observe that $\|f\|$ is smaller in case $\mathcal{P}_1$ than in the optimal case. As we discussed above, this shows that just minimizing $\|f\|$ is not sufficient to minimize $\mathcal{S}$.

Note that the choice of poles in Theorem 2.6 is quite common. In practice, we often require $\mathcal{P}$ to be in the left half plane so that the closed-loop matrix $A - bf^T$ is *stable*. Also, one often does the pole placement only on the eigenvalues of A with nonnegative real parts, while keeping the rest of the eigenvalues fixed, since this can save a lot of computational work; see [8].

We have seen under the assumptions of Theorem 2.6 that $\min \|f\|$ is achieved with poles on the line $-\rho + yi$. From (15) of Theorem 2.3, $\mathcal{S}_F \approx \|w\|_2 \|f\|_2^2$. When the pole set $\mathcal{P}$ moves away from the line $-\rho + yi$, $\|f\|$ will increase. Clearly it is possible that $\|w\|$ decreases, but $\|f\|$ enters quadratically in $\mathcal{S}$, so an increasing magnitude of $\|f\|$ usually quickly compensates a decreasing magnitude of $\|w\|$.

Although we are not able to minimize the condition number of the SIPP problem directly, our analysis of the governing factors in this condition number leads us to the following conjecture.

CONJECTURE. *Suppose that (A, b) is controllable and that A has all eigenvalues in the right half plane. If we require that $\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\} \subset \mathcal{C}^-_{-\rho}$, then the pole placement problem with minimal condition number is achieved when all elements of $\mathcal{P}$ are on the line $-\rho + yi$.*

Such a selection of poles may help to check the conditioning of the SIPP problem and may also be a good choice for the initial poles in an optimization of $\mathcal{S}$.

3. Numerical experiments. In this section we will give some further numerical examples which illustrate the theoretical results from the previous sections, and we will also illustrate other factors that contribute to the conditioning and that may be used to improve the bounds. A factors that contributes to the conditioning is the width in the set of imaginary parts of $\Lambda(A)$, but in the case of a stabilization problem, the distance between the real parts of the unstable eigenvalues of A and the desired new locations for these is also a contributing factor.

The observations made in this section are still based only on numerical experiments, but they indicate directions of research.

In all examples, the spectral condition number of the closed loop matrices κ were computed by first forming the matrix G in (5) and then applying the Matlab *cond* function. The feedback gain f was generated either via the Miminis–Paige algorithm [14] or via the formulas (5)–(7).

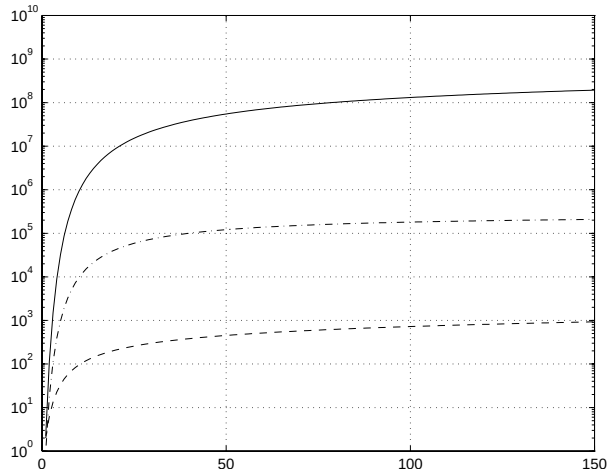


FIG. 5. *Conditioning in Example* 6: *poles reflected at the imaginary axis* $\mathcal{S}$:—, $\|f\|$: - - -, κ: $-\cdot-$.

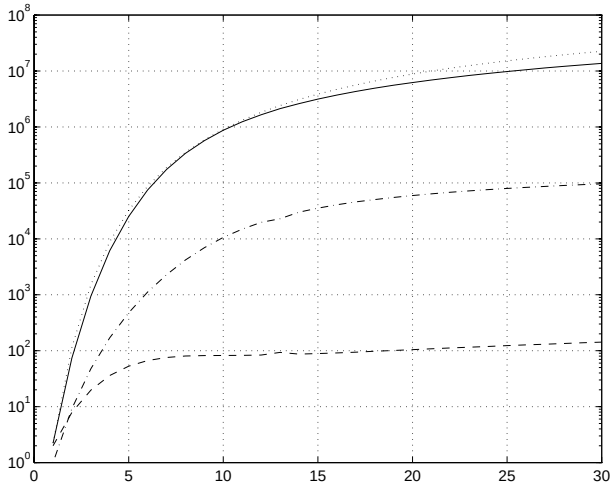


FIG. 6. *Optimal conditioning in Example* 6. $\mathcal{S}$:—, $\mathcal{S}_{ref}$: $\cdots$, $\|f\|$: - - -, κ: $-\cdot-$.

Our first example demonstrates that a certain geometric relationship between the eigenvalues of A and the chosen poles can lead to a reasonable conditioning.

Example 6. Let $A = I_n + i\,\mathrm{diag}(-\frac{n-1}{2}, -\frac{n-3}{2}, \ldots, \frac{n-3}{2}, \frac{n-1}{2})$, $b = e$, and $\mathcal{P} \subset \mathcal{C}_{-1}^{-}$.

Let $\mathcal{P}$ be the set of poles obtained by reflecting the eigenvalues of A about the imaginary axis, i.e., $\mathcal{P} = \{\lambda_1, \ldots, \lambda_n\}$ and $\lambda_j = -1 + i\frac{n+1-2j}{2}$. The values of $\|f\|$, κ, and $\mathcal{S}$ with $n = 1:150$ are displayed in Figure 5. For comparison we again used a "random search" method for $n = 1:30$ to determine a set of "optimal" poles $\mathcal{P}_0$. The condition estimates for $\mathcal{P}_0$ and $\mathcal{P}$ are given in Figure 6. In both cases, $\mathcal{S}$ increases monotonially with n, but as $n > 10$, it grows very slowly. In the first case, when $n = 150$, $\mathcal{S} = 1.9362 \times 10^8$.

The magnitude of $\mathcal{S}$ is reduced with the pole set $\mathcal{P}_0$, but just marginally. We also depict $eps \cdot \mathcal{S}$, and the error for the closed-loop eigenvalues E_n, in Figures 7 and 8 for both cases.

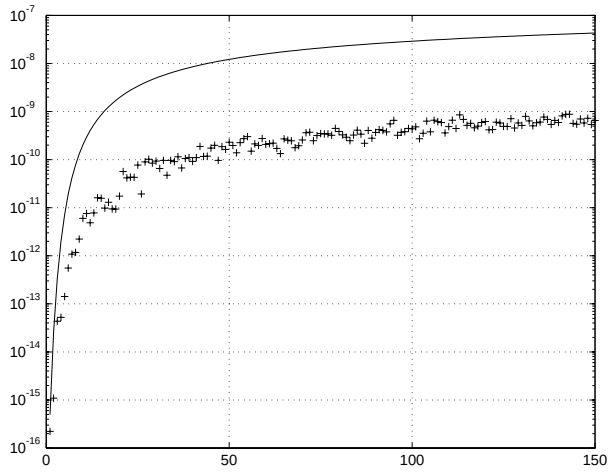


FIG. 7. *Eigenvalue error and eps · S in Example* 6 *with poles reflected at the imaginary axis.* eig_e : + + +, *eps* · $\mathcal{S}$: —.

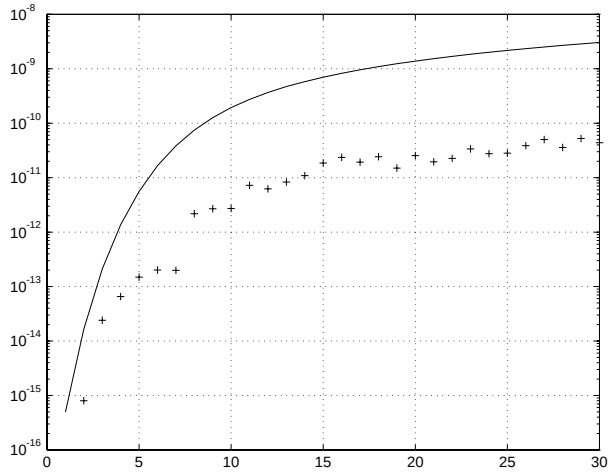


FIG. 8. *Eigenvalue error and eps · S in Example* 6 *with "optimal poles"* eig_e : + + +, *eps* · $\mathcal{S}$: —.

Unlike in Example 5, the conditioning of the SIPP problems in Example 6 seems to be reasonable. This asks for an explanation. Let $A = \rho_1 I_n + i \cdot d \cdot \mathrm{diag}(-\frac{n-1}{2}, \ldots, \frac{n-1}{2})$, $b = e$, and $\mathcal{P} = \{-\rho_2 - d\frac{n-1}{2}i, \ldots, -\rho_2 + d\frac{n-1}{2}i\}$. Suppose that $d > 0$, $\rho_1, \rho_2 > 0$. Introducing $\rho := \rho_1 + \rho_2$ we obtain, for each component of f,

$$
\begin{aligned}
|f_j|^2 &= \frac{\prod_{k=1}^{n} |\gamma_j - \lambda_k|^2}{\prod_{k=1,k\neq j}^{n} |\gamma_j - \gamma_k|^2} \\
&= \frac{\prod_{k=1}^{n} (\rho^2 + d^2(j - \frac{n+1}{2} - (k - \frac{n+1}{2}))^2)}{d^2 \prod_{k=1,k\neq j}^{n} (j - \frac{n+1}{2} - (k - \frac{n+1}{2}))^2} \\
&= \rho^2 \prod_{k=1,k\neq j}^{n} \left(1 + \frac{(\rho/d)^2}{(j-k)^2}\right)
\end{aligned}
$$

$$= \rho^2 \prod_{k=1}^{j-1} \left(1 + \left(\frac{\rho}{d}\right)^2 \frac{1}{k^2}\right) \prod_{k=1}^{n-j} \left(1 + \left(\frac{\rho}{d}\right)^2 \frac{1}{k^2}\right)$$
$$\leq \rho^2 e^{(\frac{\rho}{d})^2 \sum_{k=1}^{n-j} \frac{1}{k^2}} e^{(\frac{\rho}{d})^2 \sum_{k=1}^{j-1} \frac{1}{k^2}} \leq \rho^2 e^{4(\frac{\rho}{d})^2}.$$

So, $\|f\|_2 \leq \sqrt{n}\rho e^{2(\frac{\rho}{d})^2}$. Similarly, we get $\|w\|_2 \leq \sqrt{n}\rho e^{2(\frac{\rho}{d})^2}$ and hence

$$\mathcal{S}_F \approx \|f\|_2 \kappa_F \leq \frac{n}{\rho} \|w\|_2 \|f\|_2^2 \leq n^{\frac{5}{2}} \rho^2 e^{6(\frac{\rho}{d})^2}. \tag{20}$$

In Example 6 we have $\rho = 2$, $d = 1$, so $\mathcal{S} \leq 1.06 n^{\frac{5}{2}} \times 10^{11}$. Clearly this is a large overestimate, but the bound increases polynomially in n. Consider a generalized problem as Example 5, with $A = diag(\rho_1 + d, \ldots, \rho_1 + (n-1)d)$, $b = e$, and $\mathcal{P} = \{-\rho_2 - d, \ldots, -\rho_2 - (n-1)d\}$. Suppose also that $d, \rho_1, \rho_2 > 0$ and let $\rho := \rho_1 + \rho_2$. Then for the nth components of f and w, we have

$$|f_n| = |w_n| = (\rho + 2(n-1)d) \frac{\prod_{k=1}^{n-1} (\rho/d + (n-2+k))}{(n-1)!}$$
$$> (\rho + 2(n-1)d) \frac{(2n-3)!}{(n-1)!(n-2)!}.$$

Hence by (15),

$$\mathcal{S}_F \geq \frac{n}{(\rho + 2(n-1)d)^2} |f_n|^2 |w_n| \sim O(4^{3n}).$$

So $\mathcal{S}_F$ increases exponentially in n regardless of the magnitude of $\frac{\rho}{d}$.

We also see the importance of the scalar $\frac{\rho}{d}$ for the problems in Example 6. The smaller it is, the better conditioned the related SIPP problem will be. Since $\frac{\rho}{d}$ is determined not only by d, which describes the width in the set of imaginary parts of the eigenvalues of A, but also by the distance between the real parts of $\Lambda(A)$ and those in $\mathcal{P}$, one should observe that $\mathcal{S}$ changes when the real parts of $\Lambda(A)$ vary. Consider the following example.

Example 7. Let

$$A = \frac{1}{2} I_{31} - i \cos\left(\frac{(m-1)\pi}{180}\right) \mathrm{diag}(-15, -14, \ldots, 14, 15)$$
$$+ \sin\left(\frac{(m-1)\pi}{180}\right) \mathrm{diag}(15, 14, \ldots, 14, 15),$$

$m = 1:90$, $b = e$. $\mathcal{P}$ is selected in the following two different ways.

Case I. $\mathcal{P} = -\lambda(\bar{A})$.

Case II. $\mathcal{P} = \{-0.5 - 15\cos(\frac{(m-1)\pi}{180})i, \ldots, -0.5 + 15\cos(\frac{(m-1)\pi}{180})i\}$.

The condition estimates are depicted in Figure 9. As m increases, ρ grows and d decreases, so $\mathcal{S}$ grows, too. The test results support this observation.

We have so far mostly considered matrices A with regular eigenvalue patterns, but the same behavior is observed in the general case.

From all examples that we have tested, we see that the wider the set of imaginary parts of $\Lambda(A)$ and the smaller the distance between the set of real parts of $\Lambda(A)$ to that of $\mathcal{P}$, the better conditioning of the corresponding SIPP problem we can expect.

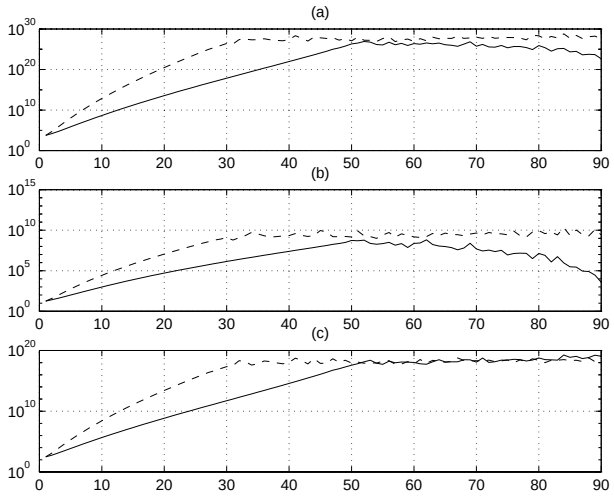


FIG. 9. *Conditioning in Example* 7: (a) $\mathcal{S}$; (b) $\|f\|$; (c) κ; *Case* I: - - -, *Case* II: —.

With these observations, the class of SIPP problems in Example 5 is probably the worst class of problems, while the class of problems in Example 6 is probably the best class.

In many control problems, for example in robust stabilization, the choice of the imaginary parts of $\Lambda(A)$ seems to be unimportant. But the above observations indicate that the imaginary parts are of great importance since they participate heavily in the conditioning of the SIPP problem.

We also tested many examples with multiple poles in $\mathcal{P}$, but then, as could be expected, the conditioning is much worse than in the case of distinct poles. If the matrix A has eigenvalues with negative real parts, which can also be moved to improve the conditioning, then it can be predicted that $\mathcal{S}$ can be reduced compared to the case when just the eigenvalues with nonnegative real parts are assigned. Numerical examples support this prediction. For more examples, see [13].

4. Conclusions. In this paper we have studied the problem of optimizing the conditioning of the single input pole placement problem when the poles are allowed to vary in specific regions of the complex plane. It is in general very difficult to minimize the condition number or the bounds for the eigenvalue error (or the error in f) in the SIPP problem, since these functions are very complicated and, in particular, for large n a numerical minimization seems prohibitive. To get arround this difficulty, we have studied two of the factors in the perturbation bounds, f and κ, the scaled spectral condition number of the closed-loop matrix and, we have neglected the effects of the condition number of the matrix A and the distance to uncontrollability of (A, b).

For problems where the optimization of f and κ can be carried out explicitly, we have determined formulas for the minima. From these formulas we are motivated to conjecture the location for the optimal pole selection in the important stabilization problem, where $\Lambda(A) \subset \mathcal{C}_0^+$, $\mathcal{P} \subset \mathcal{C}_{-\rho}^-$.

By several numerical tests we have indicated how the conditioning of the SIPP problem is determined by the distribution of the eigenvalues of A and the geometric relationship to the selected poles.

Also, we have pointed out that in order to study the accuracy of the results of the SIPP problem, it is not enough to consider just the accuracy of the feedback gain.

In Example 4, the computed gain vectors have no correct digits, while the associated eigenvalues of the computed closed-loop matrix still have about eight correct digits. But as shown by Example 1, the converse may also be the case, i.e., even though f is very accurate, the poles of the computed closed-loop system are far away from the desired poles.

Acknowledgment. We thank an anonymous referee for helpful comments which improved the presentation and readability of the paper.

REFERENCES

[1] J. Ackermann, *Der Entwurf linearer Regelungssysteme im Zustandsraum*, Regelungstechnik und Prozessdatenverarbeitung, 7 (1972), pp. 297–300.
[2] M. Arnold, *Algorithms and Conditioning for Eigenvalue Assignment*, Ph.D. thesis, Dept. of Mathematics, Northern Illinois University, De Kalb, IL, 1993.
[3] D. Boley, *Estimating the sensitivity of the algebraic structure of pencils with simple eigenvlaue estimates*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 632–643.
[4] W. L. Brogan, *Applications of a determinant identity to pole-placement and observer problems*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 612–614.
[5] J. Demmel, *The condition number of equivalence transformations that block diagonalize matrix pencils*, SIAM J. Numer. Anal., 20 (1983), pp. 599–610.
[6] T. Finck, G. Heinig, and K. Rost, *An inversion formula and fast algorithms for Cauchy-Vandermonde matrices*, Linear Algebra Appl., 183 (1993), pp. 179–191.
[7] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., Johns Hopkins Univ. Press, Baltimore, MD, 1989.
[8] C. He, A. J. Laub, and V. Mehrmann, *Placing Plenty of Poles is Pretty Preposterous*, preprint 95 − 17, DFG-Forschergruppe Scientific Parallel Computing, Fak. f. Mathematik, TU Chemnitz-Zwickau, D-09107 Chemnitz, FRG, May 1995.
[9] C. He and V. Mehrmann, *Stabilization of large linear systems*, in Proceedings IEEE Workshop on Computer-Intensive Methods in Control and Signal Processing, Prague, September 1994, M. Karny and K. Warwick, eds. IEEE Computer Society Press, Los Alamitos, CA, pp. 91–100; also see preprint 94 − 21, DFG-Forschergruppe Scientific Parallel Computing, Fak. f. Mathematik, TU Chemnitz-Zwickau, D-09107 Chemnitz, FRG, October 1994.
[10] J. Kautsky, N. K. Nichols, and P. Van Dooren, *Robust pole assignment in linear state feedback*, Internat. J. Control, 41 (1985), pp. 1129–1155.
[11] M. M. Konstantinov and P. Hr. Petkov, *Conditioning of Linear State Feedback*, Technical report 93-61, Dept. of Engineering, Leicester University, 1993.
[12] V. Mehrmann and H. Xu, *An analysis of the pole placement problem.* I. *The single-input case*, ETNA, 4 (1996), pp. 89–105.
[13] V. Mehrmann and H. Xu, *Choosing Poles So That The Single-Input Pole Placement Problem Is Well-conditioned*, preprint SFB 393/96-01, Sonderforschungsbereich 393, Numerische Simulation auf massivparallelen Rechnern, Fak. f. Mathematik, TU Chemnitz-Zwickau, D-09107 Chemnitz, FRG, January 1996.
[14] G. S. Miminis and C. C. Paige, *A direct algorithm for pole assignment of time-invariant multi-input linear systems using state feedback*, Automatica, 24 (1988), pp. 343–356.
[15] P. Hr. Petkov, N. D. Christov, and M. M. Konstantinov, *A computational algorithm for pole assignment of linear multiinput systems*, IEEE Trans. Automat. Control, AC-31 (1986), pp. 1044–1047.
[16] V. Sima, *An efficient Schur method to solve the stabilization problem*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 724–725.
[17] J. G. Sun, *Perturbation analysis of the pole assignment problem*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 313–331.
[18] A. Varga, *A Schur method for pole assignment*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 517–519.
[19] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.
[20] W. Wonham, *Linear Multivariable Control. A Geometric Approach*, Springer-Verlag, Berlin, Heidelberg, 1974.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 682–695, July 1998

005

NODE SELECTION STRATEGIES FOR BOTTOM-UP SPARSE MATRIX ORDERING*

EDWARD ROTHBERG† AND STANLEY C. EISENSTAT‡

Abstract. The minimum degree and minimum local fill algorithms are two bottom-up heuristics for reordering a sparse matrix prior to factorization. Minimum degree chooses a node of least degree to eliminate next; minimum local fill chooses a n ode whose elimination creates the least fill. Contrary to popular belief, we find that minimum local fill produces significantly better orderings than minimum degree, albeit at a greatly increased runtime. We describe two simple modifications to this strategy that further improve ordering quality. We also describe a simple modification to minimum degree, which we term approximate minimum mean local fill, that reduces factorization work by roughly 25% with only a small increase in runtime.

Key words. sparse matrices, ordering algorithms, minimum degree, minimum local fill, minimum deficiency, graph algorithms

AMS subject classifications. 65F05, 65F50

PII. S0895479896302692

1. Introduction.

When solving a symmetric, positive definite system of equations $Ax = b$ using a direct method, it is typically cheaper to factor a permuted matrix PAP^T than the original matrix A. The most commonly used heuristic for computing a permutation P that reduces fill in the factor matrix is the minimum degree algorithm [13], [17], [22]. Simply stated, minimum degree always chooses a node of least degree to eliminate next, but in effect this choice also minimizes a coarse upper bound on the amount of fill created when the node is eliminated. A less popular alternative is the minimum local fill or minimum deficiency algorithm [17], [22], which always chooses a node whose elimination creates the least amount of fill.

Here we investigate these and several other approaches to selecting nodes for elimination. Contrary to popular belief, we find that minimum local fill produces significantly better orderings than minimum degree, albeit at a greatly increased runtime. We describe two simple modifications to this heuristic that produce even better orderings, but unfortunately the runtimes are still prohibitive. Thus we also explore simple approximations to these fill metrics with the goal of reducing the runtime. The best of these approximations is no more difficult to compute than the degree, yet the orderings produced require roughly 25% less factorization work than those produced by minimum degree.

The organization of the paper is as follows. Section 2 describes the minimum degree and minimum local fill ordering heuristics, including several enhancements that have been made to the basic algorithms over the years. Section 3 describes some modifications to these methods. Section 4 presents results for all of these heuristics, and section 5 attempts to interpret them. Finally, section 6 presents conclusions.

*Received by the editors April 22, 1996; accepted for publication (in revised form) by J. W. H. Liu August 18, 1997; published electronically March 18, 1998.
http://www.siam.org/journals/simax/19-3/30269.html

†ILOG, Inc., 1901 Landings Dr., Mountain View, CA 94043 (rothberg@ilog.com). This work was done while this author was at Silicon Graphics, Inc., Mountain View, CA 94043.

‡Department of Computer Science, Yale University, P.O. Box 208285, New Haven, CT 06520-8285 (stanley.eisenstat@yale.edu). The research of this author was supported in part by NSF grant CCR-9400921.

This paper is a significant revision of an earlier paper by one of the authors [21]. Recently, Ng and Raghavan [19] have also looked at approximate fill metrics.

2. Minimum degree and minimum local fill ordering. This section briefly describes the minimum degree and minimum local fill ordering heuristics. We assume that the reader has some familiarity with the relevant concepts and algorithms. See [12], [13], [15] for more details.

Sparse matrix ordering methods are most easily described in terms of the *elimination graph* of A [20], the undirected graph $G = (V, E)$ that contains a node $j \in V$ for every column in A and an edge $e_{ij} \in E$ for every nonzero value A_{ij}, $i \neq j$. The edge $e_{ij} \in E$ is *incident* to the nodes i and j. The *degree* of a node j is the number of edges incident to j. Two nodes, i and j, are *adjacent* if $e_{ij} \in E$. The set $Adj(j)$ is the set of nodes that are adjacent to node j.

The factorization of a symmetric matrix A can be modeled by a sequence of elimination graphs, $G^k = (V^k, E^k)$, where each G^k captures the nonzero structure of the matrix that remains after k columns of A have been eliminated. Graph G^0 is the graph of A. In matrix terms, the elimination of a column causes the outer-product matrix αzz^T to be added into the remainder of A, where z is the off-diagonal part of the eliminated column. Since this matrix is symmetric, only the lower (or upper) triangle need be computed. In graph terms, the graph G^k is obtained from the graph G^{k-1} by removing from V^{k-1} the node x_k corresponding to the kth eliminated column, removing from E^{k-1} all edges incident to x_k, and adding to E^{k-1} any edges needed to make the neighbors of x_k pairwise adjacent. The added edges correspond to *fill* in the factor matrix.

As mentioned earlier, the *minimum degree* heuristic performs the ordering by eliminating at each stage k a node x_k that minimizes $|Adj_{G^{k-1}}(x_k)|$, where $Adj_{G^{k-1}}(i)$ is the set of nodes adjacent to node i in the graph G^{k-1}. The elimination of x_k makes its neighbors pairwise adjacent, thereby introducing at most $d(d-1)/2$ new edges (where $d = |Adj_{G^{k-1}}(x_k)|$). Since $d(d-1)/2$ is a strictly increasing function for $d \geq 1$, this choice of x_k is equivalent to eliminating a node that minimizes a coarse *upper bound* on the amount of fill.

The *minimum local fill* or *minimum deficiency heuristic* [22], [17] uses the exact amount of fill rather than the bound above to select a node for elimination. This approach is generally thought to provide limited quality advantages over minimum degree while requiring significantly higher runtime. To quote Duff, Erisman, and Reid [6, p. 135], "Our conclusion is that the strategies of Markowitz and minimum degree are the best local algorithms ... more complicated heuristics greatly add to the algorithm time and have little impact on performance." We should add, however, that subsequent studies [4], [16], [18], especially in connection with interior point methods for linear programming, have obtained somewhat better ordering quality from minimum local fill.

2.1. Quotient graphs. The minimum degree heuristic, while simple to describe, has proven quite difficult to implement efficiently [13]. The obvious approach, where the elimination graphs G^k are maintained as sets of edges, is inefficient in both storage and runtime. Efficient implementations require an alternative representation of G^k based on the concept of a *quotient graph* [10], [12].

The most important data structure in the quotient graph representation of G^k is the *clique*, a set of nodes that are pairwise adjacent. Recall that during the factorization the nodes in $Adj_{G^{k-1}}(x_k)$ form a clique in G^k. We use C_k to refer to both the set of nodes in this clique and the set of edges they induce; the specific meaning should be apparent from the context.

```
for k = 1 to n
    Choose a node x_k of minimum degree
    Form clique C_k = Adj(x_k)
    Destroy all cliques C_ℓ where x_k ∈ C_ℓ
    for all nodes i ∈ C_k
        Add C_k to the list of cliques to which i belongs
        Update the degree of i
```

FIG. 2.1. *Minimum degree ordering using quotient graphs.*

The edges in the graph G^k can be concisely represented as the union of the original edges E^0 and the cliques C_1 through C_k created by the elimination of nodes x_1 through x_k. More precisely, G^k is the induced graph (on the set V^k of uneliminated nodes) that is formed by discarding all nodes not in V^k and all edges incident to those nodes. Thus when we refer to a clique C_j in the graph G^k we assume that it does not contain any eliminated nodes (i.e., $C_j \subset V_k$). Note that C_j becomes redundant once any node $x_\ell \in C_j$ is eliminated since the clique C_ℓ is a superset of C_j in the induced graph.

The quotient graph representation of G^k can be shown to require no more storage than the original graph G^0 for all k [11]. Furthermore, the information required for the minimum degree algorithm can be computed and updated efficiently by simply keeping track of the set of cliques to which each uneliminated node belongs. Given this representation, for example, one can compute the set $Adj_{G^k}(i)$ of nodes adjacent to a node i in the graph G^k by computing the union of all cliques to which node i belongs and adding the original neighbors of i in G^0. The minimum degree algorithm, expressed in terms of quotient graphs, is shown in Figure 2.1. This implementation is typically referred to as quotient minimum degree (QMD).

2.2. Enhancements to minimum degree ordering. As the minimum degree algorithm has evolved, there have been several enhancements. This section briefly describes those that are relevant to the methods presented here. See [13] for more details.

Perhaps the most important enhancement is the notion of a *supervariable* [9] or *supernode* [12] in G^k. A supernode Q is a set of nodes that satisfies $Adj(i) \cup \{i\} = Adj(j) \cup \{j\}$ for all pairs of nodes $i, j \in Q$. Supernodes possess two crucial properties. The first is that all nodes in a supernode can be eliminated consecutively in a minimum degree ordering. To understand why, note that the nodes in a supernode Q all have the same degree and that eliminating one node in Q decreases the degree of every other node in Q by one. Thus, if the degree of the original node were the minimum, then the degrees of the remaining nodes become the new minimum. The second important property is that a supernode in G^k remains a supernode (or is subsumed by a larger supernode) in all subsequent graphs G^ℓ for $\ell > k$. A consequence of these two properties is that the minimum degree algorithm can treat a supernode as a single logical node. This can dramatically reduce the number of distinct nodes in the graph, resulting in significant reductions in ordering runtime.

We represent supernodes using capital letters. We define the set $Adj(I)$ of nodes adjacent to the supernode I to be $Adj(i) \backslash I$ for any $i \in I$.

A related enhancement is the use of *external degree* rather than true degree to choose the node to eliminate next [15]. The true degree of a node i in supernode I is $|Adj(i)|$; the external degree is $|Adj(I)|$. In other words, the external degree

excludes nodes in the same supernode from the degree calculation. The motivation is that if node j belongs to the same supernode as i, then j is already adjacent to the other nodes in $Adj(i)\backslash\{i\}$. Thus the only edges added by the elimination of node i are between nodes in $Adj(I)$. There are at most $d(d-1)/2$ such edges (where $d = |Adj(I)|$), and again this is minimized by minimizing $|Adj(I)|$. The use of this tighter bound on the amount of fill leads to better orderings [13], [15]. All ordering methods considered in this paper use external degrees rather than true degrees.

A third enhancement is the use of *approximate degrees* rather than exact degrees [1], [14]. Rather than computing $|Adj(i)|$ when the degree of a node i is updated, the approximate minimum degree (AMD) algorithm [1] approximates the degree using the sizes of the most recently created clique C_k and the intersections of C_k with the other cliques to which i belongs:

$$|C_k| + \sum_{j \le k, i \in C_j} |C_j \backslash C_k|.$$

This quantity is then augmented by the number of nodes in G^0 that are adjacent to i but do not belong to any clique containing i. Note that nodes that belong to more than one clique C_j are counted more than once in this expression so it provides only an upper bound on the exact degree. Nonetheless, the orderings computed using approximate degrees are of comparable quality to those obtained using exact degrees [1]. Moreover, the approximate degree is much less expensive to compute—the sizes of the sets $C_j \backslash C_k$ can be reused when updating the approximate degrees of other nodes in C_k, but a similar set subtraction computation is required each time the exact degree of a node in $C_j \cap C_k$ is updated.

A fourth enhancement is *compression* of the original graph G^0 by identifying supernodes [2], [5]. This greatly reduces the size of some initial graphs. But while it reduces the cost of the multiple minimum degree (MMD) implementation of minimum degree [15], our experience indicates that the reduction in runtime for AMD is roughly equal to the cost of performing the compression. We use it nonetheless because several of our scoring functions benefit from initial supernode information.

2.3. Minimum local fill ordering. The minimum local fill heuristic has received much less attention in the literature than minimum degree, primarily because its runtime is prohibitive. To compute the fill that would result from the elimination of a node k, we must determine which pairs of nodes in $Adj(k)$ are already adjacent, and this is much more expensive than simply computing $|Adj(k)|$. To compound the problem, while the elimination of a node k can only affect the degrees of nodes in $Adj(k)$, it can affect fill counts for both nodes in $Adj(k)$ and their neighbors. While many of the enhancements described above for minimum degree are applicable to minimum local fill (particularly supernodes), runtimes are still prohibitive.

3. New node selection strategies. This section describes several simple modifications to the minimum local fill and minimum degree algorithms that improve the quality of the computed orderings. We give an intuitive explanation of their effectiveness in this section and attempt to provide a more formal explanation in section 5.

To more easily describe these new heuristics, we introduce the function $score(K)$ that captures the "cost" of eliminating an uneliminated supernode K. The ordering algorithm always chooses a node with minimum score to eliminate next. In the case of minimum degree, $score(K) = |Adj(K)|$; for minimum local fill, $score(K) = |Fill(K)|$, where $Fill(K)$ is the set of edges that would be added if K were eliminated.

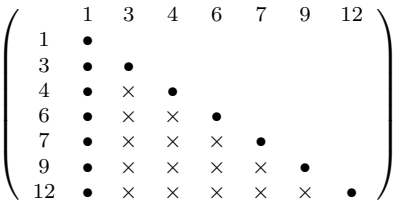


FIG. 3.1. *Potential consequences of eliminating a node adjacent to the set* {3,4,6,7,9,12}. *The symbol* × *represents possible fill-in; other symbols represent known nonzero values.*

3.1. Enhancements to minimum local fill. We consider two simple modifications to the minimum local fill node selection strategy.

The first is motivated by the observation that eliminating a supernode K corresponds to $|K|$ single-node eliminations, so the average fill associated with each elimination is $score(K) = |Fill(K)|/|K|$. We call this scoring function minimum mean local fill (MMF).

The second is based on the observation that when a supernode K is eliminated, the elimination both adds and removes edges incident to the neighbors of K. The edges added are the fill edges $Fill(K)$; the edges removed are those between K and $Adj(K)$. The net change in the number of edges is $score(K) = |Fill(K)| - |Adj(K)| \times |K|$. We call this scoring function minimum increase in neighbor degree (MIND) since it measures the aggregate change in degree of all neighbors of the eliminated node.

Note that both modifications favor the elimination of larger supernodes.

While we have given an intuitive explanation of their effectiveness, we caution the reader that the results obtained with these and similar scoring functions are often far from intuitive. Using $score(K) = |Fill(K)|/\sqrt{|K|}$, for example, actually produced slightly better results than the scoring functions described above.

3.2. Enhancements to minimum degree. We now consider a modification to the minimum degree node selection strategy. Our goal is to introduce some of the flavor of minimum local fill without also introducing the prohibitive cost.

Recall that the motivation for using degree to choose nodes for elimination is that it corresponds to a simple upper bound on fill. For example, Figure 3.1 shows the lower triangle of the submatrix of A affected by the elimination of node 1, where $Adj(1) = \{3, 4, 6, 7, 9, 12\}$. The degree of node 1 is $d = 6$, and eliminating that node could create as many as $(d^2 - d)/2 = 15$ new nonzero entries (we assume that the diagonal entries are already nonzero).

Our enhancement to the minimum degree algorithm exploits the fact that we can inexpensively identify some entries that are modified by the elimination but are already nonzero and thus cannot suffer fill. In particular, if the graph G^k contains a clique C_j, then all pairs of nodes in that clique are already adjacent. This information can be used to improve the bound on the amount of fill generated by eliminating a node. For example, Figure 3.2 shows the same submatrix as before, but with the assumption that G^k contains two cliques, $\{1, 3, 4, 6\}$ and $\{1, 4, 7, 9\}$. The bound on fill derived using only the degree is 15; but if all entries known to be nonzero due to the existence of these two cliques are excluded, the bound is reduced to 9.

In our approximation we define $score(I) = (d^2 - d)/2 - (c^2 - c)/2$, where d is the external degree of the supernode I (either exact or approximate) and $c = |C_\ell \backslash I|$, where C_ℓ is the most recently created clique containing I. Note that this score need

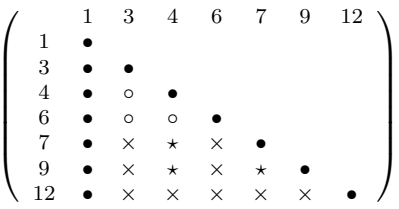


FIG. 3.2. *Potential consequences of eliminating a node adjacent to the set* {3,4,6,7,9,12}, *assuming A contains two cliques,* $\{1,3,4,6\}$ *and* $\{1,4,7,9\}$. *The symbol* $\times$ *represents possible fill-in; other symbols represent known nonzero values.*

only be updated when a neighbor x_k of I is eliminated, in which case ℓ will be equal to k. Moreover, it is trivial to compute since any implementation of minimum degree already computes all of the relevant quantities. In the example of Figure 3.2, the external degree of node 1 is $d = 6$, the size of the clique C_ℓ is $c = 3$ (using either clique), and thus $score(1) = 12$. We call this approach approximate minimum local fill (AMF).

This simple bound on fill can often be tightened by using the largest clique to which supernode I belongs rather than the most recently created one [21]. It can be further tightened by taking into account the effects of multiple cliques [19], [21]. Considering both cliques, the previous example gives a bound of 9 versus the bound of 12 obtained by considering a single clique. However, we found that tightening the fill bound beyond the information available from the most recently formed clique did not improve ordering quality [21].

We also consider approximate minimum mean local fill (AMMF) and approximate minimum increase in neighbor degree (AMIND). The scoring functions for these variants are obtained by modifying the base score above in the same way that the base score $|Fill(K)|$ was modified for minimum local fill; i.e., $score(K) = score_{AMF}(K)/|K|$ for AMMF and $score(K) = score_{AMF}(K) - |deg(K)| \times |K|$ for AMIND, where $score_{AMF}(K)$ is the AMF score for supernode K and $deg(K)$ is the external degree of K (either exact or approximate).

4. Results. To evaluate the effectiveness of these scoring functions, we look at ordering quality over a set of 40 sparse symmetric matrices, including 21 matrices from the Harwell–Boeing sparse matrix test set [7] and 19 structural analysis and computational fluid dynamics matrices, 11 from NASA and 8 extracted from commercial applications. Each of these matrices require more than 100 million floating-point operations to factor using Liu's MMD ordering heuristic [15]. Floating-point operation counts, our primary evaluation metric in this section, are an extremely poor predictor of runtime for problems smaller than this.

To reduce the effect of tie-breaking strategies, all nonzero and operation counts were obtained by ordering each matrix several times (randomly permuting the rows and columns before each ordering) and taking the median. For the minimum fill variants, we take the median over three permutations; for the less costly approximate minimum fill variants, we take the median over eleven permutations.

Table 4.1 shows the number of rows in each matrix, the number of nonzero values in the lower triangle of A and the number of nonzero values in the lower triangle of L and the number of floating-point operations required to perform the factorization after applying the MMD ordering.

All of our ordering results come from three codes: Liu's implementation of MMD,

TABLE 4.1
Statistics about test matrices.

Matrix	Rows	NZ in A (10^3)	NZ in L (10^3)	Operations to factor (10^6)
BCSSTK15	3948	60	658	169
BCSSTK16	4884	147	737	144
BCSSTK17	10974	219	1117	191
BCSSTK18	11948	80	645	132
BCSSTK23	3134	24	452	140
BCSSTK25	15439	133	1506	328
BCSSTK29	13992	316	1757	434
BCSSTK30	28924	1036	3835	926
BCSSTK31	35588	608	5304	2523
BCSSTK32	44609	1029	5199	1075
BCSSTK33	8738	300	2635	1308
BCSSTK35	30237	740	2760	399
BCSSTK36	23052	583	2761	618
BCSSTK37	25503	583	2819	542
BCSSTK38	8032	181	747	121
BCSSTK39	46772	1068	7560	2147
MSC10848	10848	620	2022	562
MSC23052	23052	588	2759	612
CRYSTK01	4875	160	1055	317
CRYSTK02	13965	491	5858	3940
CRYSTK03	24696	887	13431	12229
FLAP	51537	531	5558	1875
FORD2	100196	322	2427	300
PWT	36519	181	1744	216
SPHERE6	16386	65	809	140
BIKKER2	173160	514	60899	153705
COPTER2	55476	407	14137	12560
TROLL	213453	6099	99316	189687
HSCT16K	16146	515	2913	929
SRB55K	54870	1362	12402	5040
HSCT88K	88404	2001	17874	8645
FORD263K	263096	6530	37248	21723
3DTUBE	45330	1629	30860	40119
GEARBOX	153746	4617	52798	57360
STRUCT1	46949	1164	5023	1253
STRUCT2	73752	1835	9922	3955
STRUCT3	53570	613	5297	1216
STRUCT4	4350	121	2264	1798
CFD1	70656	949	39647	47196
CFD2	123440	1605	87134	169861

our implementation of the exact minimum local fill variants, and our modification of Amestoy, Davis, and Duff's implementation[1] of AMD to perform the AMF variants. We used approximate degrees for all of the approximate minimum fill results but have verified that exact degrees produce nearly identical results. We have also verified that our codes produce orderings of comparable quality to MMD when we use the AMD scoring function.

Table 4.2 shows the number of nonzero values in the factor matrix L and Table 4.3 shows the number of floating-point operations required to factor A after applying the various ordering methods, all relative to the corresponding results for MMD. The last two lines in each table give geometric means and medians over the entire set of matrices.

[1] http://www.netlib.org/linalg/amd/amdbar.f

TABLE 4.2
Nonzero values in factor matrix L (relative to MMD).

	Exact fill			Exact fill in C_k			Approximate fill		
Matrix	MF	MMF	MIND	MF	MMF	MIND	AMF	AMMF	AMIND
BCSSTK15	.94	.87	.88	.93	.88	.88	.95	.89	.89
BCSSTK16	.86	.89	.84	.87	.93	.85	.93	.97	1.00
BCSSTK17	.82	.85	.79	.91	.90	.88	.96	.90	.97
BCSSTK18	.86	.81	.83	.90	.83	.86	.91	.88	.89
BCSSTK23	.87	.83	.81	.94	.88	.85	.89	.85	.86
BCSSTK25	.85	.79	.80	.87	.79	.82	.93	.83	.87
BCSSTK29	.89	.88	.84	.93	.92	.87	.92	.98	.93
BCSSTK30	.95	.84	.85	1.04	.86	.88	1.00	.90	.91
BCSSTK31	.89	.78	.81	.94	.79	.84	.95	.83	.88
BCSSTK32	.92	.89	.88	.95	.91	.91	.97	.93	.99
BCSSTK33	.85	.86	.81	.92	.88	.85	.93	.91	.93
BCSSTK35	.96	.93	.92	.97	.95	.94	.98	.97	1.02
BCSSTK36	.93	.90	.89	.97	.94	.94	.97	.96	1.03
BCSSTK37	.94	.93	.90	.97	.95	.93	.98	.97	1.01
BCSSTK38	.92	.88	.89	.91	.91	.91	.99	.94	.95
BCSSTK39	.86	.85	.84	.98	.93	.93	.98	.96	.96
MSC10848	.89	.85	.83	.93	.92	.87	.99	1.08	.95
MSC23052	.93	.91	.90	.97	.94	.92	.97	.97	1.02
CRYSTK01	.85	.90	.84	.90	.94	.89	.92	.95	.95
CRYSTK02	.77	.79	.73	.85	.85	.85	.89	.90	.92
CRYSTK03	.72	.77	.72	.84	.85	.85	.89	.90	.90
FLAP	.73	.77	.70	.86	.87	.84	.89	.92	.93
FORD2	.88	.86	.85	.91	.90	.88	.94	.93	.95
PWT	.80	.78	.80	.97	.92	.93	.98	.96	.98
SPHERE6	.96	.89	.91	.99	.92	.94	.98	.95	.90
BIKKER2	.81	.68	.77	.85	.71	.80	.88	.72	.78
COPTER2	.83	.69	.76	.83	.72	.78	.86	.74	.80
TROLL	.76	.64	.71	.90	.67	.76	.88	.68	.79
HSCT16K	.90	.98	.87	.91	1.01	.88	.84	1.00	.84
SRB55K	.82	.80	.77	.89	.83	.84	.94	.84	.89
HSCT88K	.97	.92	.92	.98	.93	.93	.98	.94	.94
FORD263K	.95	.87	.88	.98	.89	.91	.96	.94	.96
3DTUBE	.69	.76	.63	.83	.85	.81	.89	.90	.89
GEARBOX	.79	.80	.72	.83	.84	.83	.90	.91	.92
STRUCT1	.96	.92	.92	.98	.95	.95	.99	.98	.99
STRUCT2	.90	.88	.88	.96	.95	.93	.97	.96	.98
STRUCT3	.84	.89	.81	.94	.91	.90	.96	.94	.96
STRUCT4	.75	.78	.73	.83	.78	.80	.82	.75	.79
CFD1	.60	.59	.58	.79	.69	.76	.85	.74	.78
CFD2	.54	.54	.51	.80	.73	.76	.84	.76	.81
G. Mean	.84	.82	.80	.91	.87	.87	.93	.90	.91
Median	.86	.85	.83	.92	.90	.88	.94	.93	.93

Note that the table shows results for the exact fill methods, the approximate fill methods, and methods that compute exact fill only for the neighbors of the eliminated node (i.e., the nodes in C_k). Recall that maintaining exact fill information requires updating the scores of the neighbors of the nodes in C_k as well. Since the approximate fill variants only update the scores for nodes in C_k, computing exact fill on these nodes gives an upper bound on the improvement that can be obtained by refining our approximations.

We note the following from the results:

1. Minimum local fill (MF) provides significantly better orderings than MMD. On average, it reduces nonzero values in the factor by roughly 15% and floating-point operations by roughly 30%.

TABLE 4.3
Floating-point operations to factor A (relative to MMD).

	Exact fill			Exact fill in C_k			Approximate fill		
Matrix	MF	MMF	MIND	MF	MMF	MIND	AMF	AMMF	AMIND
BCSSTK15	.91	.76	.77	.87	.74	.77	.91	.71	.79
BCSSTK16	.73	.74	.67	.75	.84	.70	.85	.91	.96
BCSSTK17	.62	.66	.57	.81	.76	.75	.90	.73	.90
BCSSTK18	.70	.58	.65	.78	.64	.69	.79	.73	.75
BCSSTK23	.73	.64	.65	.86	.73	.70	.77	.66	.71
BCSSTK25	.69	.57	.60	.74	.57	.63	.83	.63	.68
BCSSTK29	.75	.70	.63	.83	.80	.70	.82	.93	.77
BCSSTK30	.87	.64	.65	1.07	.69	.72	.99	.71	.77
BCSSTK31	.77	.55	.63	.85	.55	.66	.89	.64	.73
BCSSTK32	.82	.74	.73	.88	.78	.79	.93	.83	.90
BCSSTK33	.72	.67	.63	.86	.70	.69	.85	.73	.83
BCSSTK35	.89	.81	.81	.93	.87	.85	.96	.90	.99
BCSSTK36	.85	.74	.75	.95	.85	.86	.93	.88	.97
BCSSTK37	.90	.84	.79	.96	.86	.86	.98	.89	.97
BCSSTK38	.81	.71	.74	.79	.76	.78	.97	.83	.83
BCSSTK39	.72	.69	.66	.94	.82	.83	.96	.87	.87
MSC10848	.75	.68	.64	.84	.81	.72	.94	1.19	.87
MSC23052	.86	.79	.75	.93	.82	.80	.91	.89	.92
CRYSTK01	.72	.77	.69	.80	.84	.80	.84	.83	.88
CRYSTK02	.60	.60	.51	.70	.71	.73	.80	.77	.85
CRYSTK03	.50	.58	.50	.70	.68	.71	.79	.75	.79
FLAP	.45	.50	.40	.69	.67	.65	.75	.78	.80
FORD2	.69	.66	.64	.73	.72	.67	.82	.78	.77
PWT	.62	.57	.62	.92	.81	.85	.96	.89	.94
SPHERE6	.91	.75	.80	.96	.81	.85	.95	.86	.74
BIKKER2	.66	.46	.61	.72	.52	.64	.77	.50	.61
COPTER2	.70	.46	.56	.69	.51	.61	.75	.54	.64
TROLL	.53	.37	.48	.72	.42	.52	.76	.43	.57
HSCT16K	.80	.83	.73	.81	.87	.76	.68	.86	.65
SRB55K	.65	.60	.56	.78	.63	.68	.88	.65	.75
HSCT88K	.92	.78	.80	.93	.80	.80	.94	.83	.78
FORD263K	.85	.64	.68	.97	.71	.77	.85	.81	.82
3DTUBE	.48	.55	.40	.69	.75	.66	.78	.78	.81
GEARBOX	.63	.59	.46	.65	.63	.65	.78	.77	.79
STRUCT1	.90	.80	.80	.96	.87	.86	.99	.91	.96
STRUCT2	.75	.69	.72	.90	.84	.80	.91	.87	.91
STRUCT3	.69	.77	.62	.89	.76	.78	.94	.82	.91
STRUCT4	.56	.61	.55	.69	.60	.66	.67	.54	.63
CFD1	.37	.34	.34	.62	.44	.56	.71	.50	.60
CFD2	.32	.32	.28	.68	.53	.59	.73	.58	.67
G. Mean	.69	.63	.61	.81	.71	.72	.85	.75	.79
Median	.72	.66	.64	.83	.75	.72	.85	.78	.80

2. MMF and MIND produce even better orderings than MF.

3. AMF also computes significantly better orderings than MMD, with roughly 7% fewer nonzero values in the factor and roughly 15% fewer floating-point operations.

4. AMMF and AMIND further improve on AMF, providing roughly 25% and 20% reductions in floating-point operation counts, respectively.

5. MMF is slightly less effective than MIND, while AMMF is slightly more effective than AMIND.

6. Our approximate fill variants capture most of the benefit of computing exact fill information on nodes in C_k.

7. The methods considered here provide better orderings than MMD for almost every matrix in our test set.

4.1. Runtimes. The benefit of a better ordering must, of course, be traded off against any increase in the cost of computing that ordering. But before discussing runtime, we must first discuss two implementation details.

Minimum degree codes typically maintain a set of buckets, where bucket i contains a list of nodes of degree i. Since degrees are bounded by the number of nodes in the graph, this bucket data structure is quite compact. However, this approach does not work for minimum fill since, in the worst case, fill is proportional to the square of the number of nodes in the graph. In our implementations we keep individual buckets for nodes with scores less than the number of nodes and a single bucket for all nodes with larger scores. In the unlikely event that the best score is greater than or equal to the number of nodes in the graph, we perform a linear search through this last bucket to find a node with the lowest score [8].

Recall that the MMF and AMMF scoring functions can compute noninteger scores, which are incompatible with the use of integer buckets as above. However, we have found that discarding the fractional portions gives roughly the same results, and this is done in our implementations.

Using this bucket data structure, we found that AMF, AMMF, and AMIND added roughly 10%, 23%, and 15% to the runtime of AMD, respectively. When we looked at these runtimes in more detail, we found that the source of the increase was not the cost of computing the scoring function—a code that computed any of these other scores but then used the AMD score was always less than 2% slower. Instead, the new scoring functions are slower because they consistently compute *more* scores than AMD. This has to do with the way in which the new methods grow cliques, as will be discussed in the next section.

5. Towards a better understanding of the results. This section looks at the experimental results in more detail. We discuss our conjectures about why these methods work and identify a number of aspects that we do not understand.

Before beginning, we note that understanding the behavior of a bottom-up ordering method is an extremely difficult task. Decisions made early in the ordering process based on local information (e.g., degree) can lead to poor behavior much later on, and it is virtually impossible to trace such behavior back to any one specific decision. We are, therefore, limited to making general observations.

Recall that the clique C_k formed by the elimination of node x_k subsumes all cliques to which x_k belongs. We say that a node x_i is *interior to* C_k if the clique C_i created by the elimination of x_i is subsumed by C_k, or by some clique subsumed by C_k, and so on. The set of nodes interior to a clique forms a subgraph; the nodes in the clique form the *boundary* of that subgraph.

We believe that the main problem with minimum degree is that the cliques created in the elimination process often have nonsmooth boundaries. Theoretical support is provided by Berman and Schnitger [3], who have shown that nonsmooth boundaries can lead to asymptotically suboptimal orderings. Empirical evidence is provided in Figure 5.1, which contains two views of a scatter plot of the number of nodes in each clique C_k versus the number of nodes interior to that clique for matrix BCSSTK15. Note that AMMF generates significantly smaller cliques than AMD for a given number of interior nodes, which means that its cliques have smoother boundaries.

We believe that the alternative scoring functions produce smoother clique boundaries because of the way they form large cliques. Recall that the approximate fill scoring functions are more willing to select nodes that already belong to large cliques. As a result these variants tend to grow large cliques into larger ones. In contrast AMD forms large cliques by merging smaller ones. Empirical evidence for this conjecture is

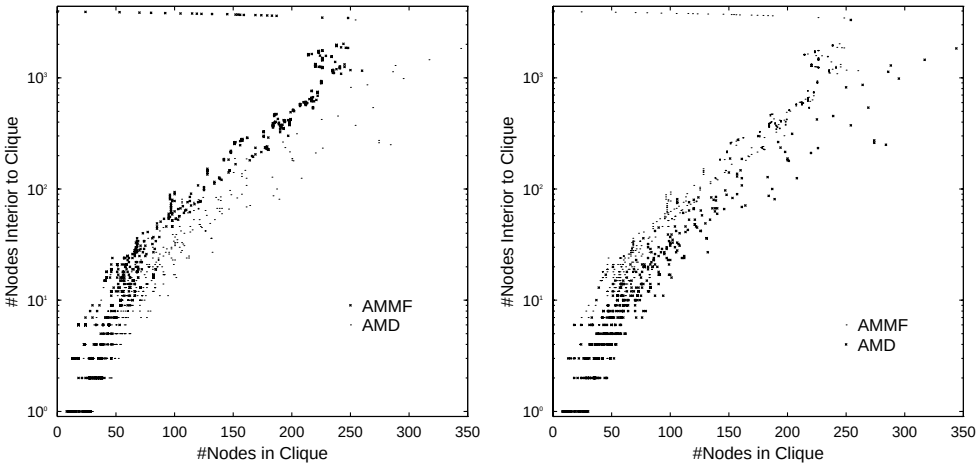


FIG. 5.1. *Number of interior nodes versus clique size for matrix* BCSSTK15.

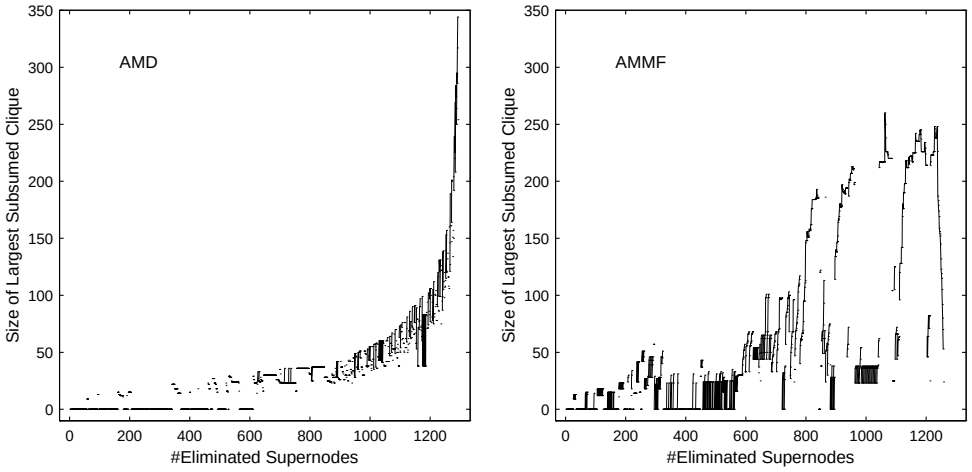


FIG. 5.2. *Largest incident clique for* BCSSTK15.

shown in Figure 5.2, which shows, for each elimination step k, the size of the largest clique to which node x_k belongs. (Lines are drawn between successive points if the clique created is immediately subsumed.) The AMMF approach exhibits significant local growth in clique sizes; in contrast clique sizes in AMD grow more smoothly.

Why should growing cliques create smoother boundaries than merging cliques? We conjecture that one very important issue is clique *alignment*. We say that two cliques are "well aligned" if they share many nodes. In general, merging poorly aligned cliques places fewer nodes in the interior of the resulting clique than merging well-aligned cliques. Intuitively, alignment is less of an issue when merging a small clique into a larger one.

Why is AMMF more effective than AMF? We found that AMMF generally grows a clique further than AMF. This is understandable since growing a clique often creates supernodes within the current clique. Since these supernodes have reduced scores in AMMF, the clique continues to grow. Apparently, growing larger cliques than those grown by AMF is beneficial. We experimented with scoring functions that encouraged cliques to continue growing beyond the point where they would stop with AMMF,

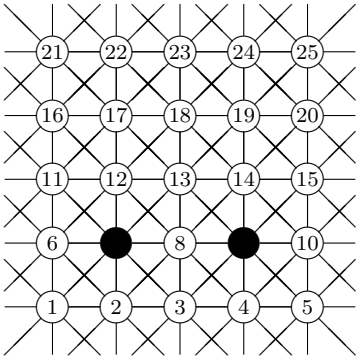


FIG. 5.3. *A portion of a* 9*-point toroidal grid.*

but the resulting orderings were worse. Taken to the extreme, of course, encouraging clique growth leads to a wavefront ordering. While growing cliques appears to be an important issue, knowing when to stop appears to be an equally important one.

When we looked at clique growth patterns for the exact fill variants, we found that they actually grew cliques less than did the approximate fill variants. Clearly, they are using a different mechanism to compute good orderings. We believe one important property that the exact fill scores capture is clique alignment. Consider the example in Figure 5.3, which shows a portion of a 9-point toroidal grid. We assume that nodes 7 and 9 have already been eliminated. The two cliques in the graph are thus $\{1, 2, 3, 6, 8, 11, 12, 13\}$ and $\{3, 4, 5, 8, 10, 13, 14, 15\}$. Minimum degree can choose any node not adjacent to the eliminated nodes to eliminate next. The method is free to choose node 18, for example. After several eliminations, this creates a patchwork of cliques. By contrast minimum fill prefers nodes 17 and 19 (and their symmetrical equivalents) whose fill count is one lower than that of node 18. Thus the graph that remains once one-fourth of the nodes have been eliminated is still a regular grid. While we cannot expect the same behavior for less regular problems, it is clear that minimum fill prefers to eliminate nodes whose elimination creates cliques that share sets of nodes with existing cliques (i.e., well-aligned cliques).

Returning briefly to the question of why runtimes for the approximate fill methods are larger than those for AMD, we consider a simplified example. Imagine you wish to build a set of n objects by merging subsets of these objects, starting with subsets of size 1. Assume that when two sets are merged, each member of the resulting set must be rescored. If you grow the set of size n by merging sets of size 1 in the first stage, sets of size 2 in the second stage, sets of size 4 in the third stage, and so on, then the total number of nodes rescored is $O(n \log n)$. If you grow the set by merging a set of size 1 into a set of size i in stage i, then the cost is $O(n^2)$. Given the fact that the approximate fill metrics favor merging small cliques into large cliques, the observed increase in node rescoring is not surprising.

To summarize, we conjecture that AMF is more effective than AMD because the process of growing cliques creates smoother clique boundaries than the process of merging smaller cliques. AMMF is even more effective because it allows the clique-growing process to continue longer. MF is still more effective because cliques must eventually be merged and exact fill scores capture some notion of clique alignment, which leads to smoother clique boundaries.

Our intent in this section has been to provide a partial explanation of some of the observed results. While we believe we understand some of the relevant issues, we feel that we have left more questions unanswered than answered. We are quite confident

that better scoring functions can be found as understanding of the underlying behavior of the methods improves.

Recently, Ng and Raghavan have also considered approximate fill metrics and obtained similar results [19]. Their fill approximation, which is similar to our AMF heuristic, uses clique information to identify edges already present in G_k. Major differences between the results are: (i) their fill approximation requires a computation similar to computing exact degrees for each node, so we would not expect runtime reductions from using approximate degrees; and (ii) for the 24 matrices in both test suites, the ordering quality for their approach lies between that of our AMF and AMIND approaches and is significantly worse than that of our AMMF approach.

6. Conclusions. We have described several simple modifications to the minimum local fill and minimum degree ordering heuristics that exploit readily available information about node adjacencies to improve the fill bounds used to select a node for elimination. Perhaps the most practical of these modifications, which we call AMMF, reduces floating-point operation counts for the subsequent factorization by roughly 25% while increasing ordering runtimes by only 23%.

Acknowledgments. The authors would like to thank Cleve Ashcraft, Tim Davis, Joseph Liu, and Farzin Shakib for their suggestions for improving the paper.

REFERENCES

[1] P. R. Amestoy, T. A. Davis, and I. S. Duff, *An approximate minimum degree ordering algorithm*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 886–905.
[2] C. Ashcraft, *Compressed graphs and the minimum degree algorithm*, SIAM J. Sci. Comput., 16 (1995), pp. 1404–1411.
[3] P. Berman and G. Schnitger, *On the performance of the minimum degree algorithm for Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 83–88.
[4] I. A. Cavers, *Using Deficiency Measure for Tiebreaking the Minimum Degree Algorithm*, Tech. report 89-2, Department of Computer Science, University of British Columbia, Vancouver, B.C., January 1989.
[5] A. C. Damhaug, *Sparse Solution of Finite Element Equations*, Ph.D. thesis, Department of Structural Engineering, Norwegian Institute of Technology, Trondheim, Norway, 1992.
[6] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*, Clarendon Press, Oxford, 1986.
[7] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.
[8] 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.
[9] I. S. Duff and J. K. Reid, *The multifrontal solution of indefinite sparse symmetric linear equations*, ACM Trans. Math. Software, 9 (1983), pp. 302–325.
[10] 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.
[11] A. George and J. W. H. Liu, *A minimal storage implementation of the minimum degree algorithm*, SIAM J. Numer. Anal., 17 (1980), pp. 282–299.
[12] A. George and J. W. H. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice–Hall, New York, 1981.
[13] A. George and J. W. H. Liu, *The evolution of the minimum degree ordering algorithm*, SIAM Rev., 31 (1989), pp. 1–19.
[14] 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.
[15] J. W. H. Liu, *Modification of the minimum degree algorithm by multiple elimination*, ACM Trans. Math. Software, 11 (1985), pp. 141–153.
[16] I. J. Lustig, R. E. Marsten, and D. F. Shanno, *Interior point methods for linear programming: Computational state of the art*, ORSA J. Comput., 6 (1994), pp. 1–14.
[17] H. M. Markowitz, *The elimination form of the inverse and its application to linear programming*, Management Sci., 3 (1957), pp. 255–269.

[18] C. Meszaros, *The Inexact Minimum Local Fill-In Ordering Algorithm*, Tech. report WP 95 7, Computer and Automation Research Institute, Hungarian Academy of Sciences, Budapest, 1995.

[19] E. G. Ng and P. Raghavan, *Performance of Greedy Ordering Heuristics for Sparse Cholesky Factorization*, manuscript, Department of Computer Science, University of Tennessee at Knoxville, 1997.

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

[21] E. Rothberg, *Ordering Sparse Matrices using Approximate Minimum Local Fill*, manuscript, Silicon Graphics, Inc., Mountain View, CA, May 1996.

[22] 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.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 696–700, July 1998

006

PERTURBATION RESULTS FOR PROJECTING A POINT ONTO A LINEAR MANIFOLD*

JIU DING†

Abstract. Some new results will be presented on the perturbation analysis for the orthogonal projection of a point onto a linear manifold. The obtained perturbation upper bound is with respect to the distance from the perturbed solution to the unperturbed manifold.

Key words. projection, least squares, generalized inverse

AMS subject classifications. 15A09, 65F20

PII. S0895479896306604

1. Introduction.

Let $A \in R^{m\times n}$ be an $m \times n$ matrix, and let $b \in R^m$ be an m dimensional vector. We consider the following perturbation problem of the orthogonal projection of a point onto a linear manifold:

$$\min \|p - x\|, \text{ subject to } \|Ax - b\| = \min_{z\in R^n} \|Az - b\|, \tag{1}$$

where $p \in R^n$ is a fixed point and the norm is the usual Euclidean 2-norm. The collection of all vectors x satisfying the constraint in (1) will be called the feasible set and its elements will be called feasible solutions of (1).

Solving the problem (1) is important in many applications, which include the usual minimal norm least squares problem ($p = 0$) and interior point methods ($b = 0$) in which the main work is the projection of the negative gradient of some so-called potential function onto the null space of some scaled matrix (see, e.g., [8]).

The unique optimal solution to (1) is given by

$$x = x(A, b, p) = A^\dagger b + (I - A^\dagger A)p,$$

where $A^\dagger$ is the Moore–Penrose generalized inverse of A; see, e.g., Theorem 3.6.2 of [3]. Thus, we have a useful expression

$$p - x = A^\dagger(Ap - b) \tag{2}$$

for the difference $p - x$ of the point p and its projection x. Such a relation between a point and its projection will be used several times in the remainder of this paper.

Since the above expression of the solution x to (1) involves the generalized inverse $A^\dagger$ of A, from the well-known discontinuity property of $A \to A^\dagger$, it is obvious that x is discontinuous at (A, b, p) with A rank-deficient. When the perturbation is rank-preserving, error estimates have been given in [1], [2], and [7] in the special case that $b \in R(A)$, and in [4] in the general case.

When the perturbation is arbitrary, that is, when it may be rank-increasing, the upper continuity property of the projection and a corresponding error estimate were obtained in [4]. However, because of a special and tedious construction, this upper bound contains an unusual "relative condition number" $\tilde{\kappa} = \|A\|\|B^\dagger\|$ of a full-ranked

*Received by the editors August 2, 1996; accepted for publication (in revised form) by L. Elden May 29, 1997; published electronically March 18, 1998.
http://www.siam.org/journals/simax/19-3/30660.html
†Department of Mathematics, The University of Southern Mississippi, Hattiesburg, MS 39406-5045 (jiu_ding@bull.cc.usm.edu).

submatrix B of A with respect to A, where $\|A\|$ is the induced Euclidean matrix norm, and a general relation between $\tilde{\kappa}$ and the condition number $\kappa = \|A\|\|A^\dagger\|$ of A is unknown. In this paper, we shall get a better error bound without using the relative condition number approach. It seems that this bound is optimal since it is with regard to the minimal distance of the perturbed solution to the manifold of the unperturbed problem, and it also looks beautiful in format since it will be reduced to well-known results in special cases. The main result in the paper has two parts. Part one covers consistent linear systems and part two covers the general systems, and they will be presented in sections 2 and 3, respectively.

2. Error bound: $b \in R(A)$. Suppose the problem (1) is perturbed to

$$\begin{aligned} &\min \quad \|(p+q)-y\|, \\ &\text{subject to} \|(A+E)y-(b+e)\| = \min_{z\in R^n} \|(A+E)z-(b+e)\|. \end{aligned} \tag{3}$$

In the following, the optimal solutions to (1) and (3) will be denoted by x^* and y^*, respectively. In this section, we assume that both linear systems $Ax = b$ and $(A+E)y = b+e$ are consistent. We always assume that $x \neq 0$ whenever $\|x\|$ appears in the denominator.

THEOREM 2.1. *Suppose $b \in R(A)$ and $b+e \in R(A+E)$. If $\|A^\dagger E\| < 1$, then there is a feasible solution x to* (1) *such that*

$$\frac{\|y^*-x\|}{\|x\|} \leq \frac{\kappa}{1-\|A^\dagger E\|}\left(\frac{\|e\|}{\|b\|}+\frac{\|E\|}{\|A\|}\right). \tag{4}$$

If, in addition, $\operatorname{Rank}(A+E) = \operatorname{Rank}(A)$ *and $\|A^\dagger\|\|E\| < 1$, then*

$$\begin{aligned} \frac{\|y^*-x^*\|}{\|x^*\|} \leq \frac{\|q\|}{\|x^*\|} + \kappa\Bigg[&\left(\frac{\|p\|}{\|b\|}+\frac{1+\sqrt{5}}{2}\frac{\|A^\dagger\|}{1-\|A^\dagger\|\|E\|}\right)\|E\| \\ &+ \frac{\|e\|}{(1-\|A^\dagger\|\|E\|)\|b\|}\Bigg]. \end{aligned} \tag{5}$$

Proof. Let x be the orthogonal projection of y^* onto the feasible set of (1). Then, replacing p with y^* in (2), we get

$$y^*-x = A^\dagger(Ay^*-b) = A^\dagger(e-Ey^*),$$

since $(A+E)y^* = b+e$. Hence,

$$(I+A^\dagger E)(y^*-x) = A^\dagger(e-Ex).$$

Now

$$0 < (1-\|A^\dagger E\|)\|y^*-x\| \leq \|(I+A^\dagger E)(y^*-x)\|$$

by the assumption. Therefore,

$$\begin{aligned} \frac{\|y^*-x\|}{\|x\|} &\leq \frac{1}{1-\|A^\dagger E\|}\|A^\dagger\|\frac{\|e-Ex\|}{\|x\|} \\ &\leq \frac{1}{1-\|A^\dagger E\|}\|A^\dagger\|\frac{\|e\|+\|Ex\|}{\|x\|} \\ &= \frac{\kappa}{1-\|A^\dagger E\|}\frac{\|e\|+\|Ex\|}{\|A\|\|x\|} \\ &\leq \frac{\kappa}{1-\|A^\dagger E\|}\left(\frac{\|e\|}{\|b\|}+\frac{\|E\|}{\|A\|}\right). \end{aligned}$$

This proves (4). To prove (5), subtracting the equality

$$p+q-y^* = (A+E)^\dagger\left[(A+E)(p+q)-(b+e)\right]$$

from the equality $p-x^* = A^\dagger(Ap-b)$, we have

$$\begin{aligned} y^*-x^* &= q+A^\dagger Ap-(A+E)^\dagger(A+E)(p+q)+(A+E)^\dagger(b+e)-A^\dagger b \\ &= [I-(A+E)^\dagger(A+E)]q+[A^\dagger A-(A+E)^\dagger(A+E)]p \\ &\quad +[(A+E)^\dagger-A^\dagger]b+(A+E)^\dagger e. \end{aligned}$$

Since $I-(A+E)^\dagger(A+E)$ is an orthogonal projector,

$$\begin{aligned} \|y^*-x^*\| &\le \|q\|+\|(A+E)^\dagger(A+E)-A^\dagger A\|\|p\| \\ &\quad +\|(A+E)^\dagger-A^\dagger\|\|b\|+\|(A+E)^\dagger\|\|e\|. \end{aligned} \tag{6}$$

Now, since $\mathrm{Rank}(A+E)=\mathrm{Rank}(A)$, from [6],

$$\|(A+E)^\dagger(A+E)-A^\dagger A\| \le \|A^\dagger\|\|E\|,$$

$$\|(A+E)^\dagger-A^\dagger\| \le \frac{1+\sqrt{5}}{2}\|(A+E)^\dagger\|\|A^\dagger\|\|E\|,$$

and

$$\|(A+E)^\dagger\| \le \frac{\|A^\dagger\|}{1-\|A^\dagger\|\|E\|}.$$

Thus, by (6), and noting that $\|A\|\|x^*\| \ge \|Ax^*\| = \|b\|$, we have

$$\begin{aligned} \frac{\|y^*-x^*\|}{\|x^*\|} &\le \frac{\|q\|}{\|x^*\|}+\frac{\|A^\dagger\|\|E\|\|p\|}{\|x^*\|}+\frac{1+\sqrt{5}}{2}\frac{\|A^\dagger\|^2\|E\|\|b\|}{(1-\|A^\dagger\|\|E\|)\|x^*\|} \\ &\quad +\frac{\|A^\dagger\|\|e\|}{(1-\|A^\dagger\|\|E\|)\|x^*\|} \\ &\le \frac{\|q\|}{\|x^*\|}+\frac{\kappa\|p\|\|E\|}{\|A\|\|x^*\|}+\frac{1+\sqrt{5}}{2}\frac{\kappa\|A^\dagger\|\|E\|\|b\|}{(1-\|A^\dagger\|\|E\|)\|A\|\|x^*\|} \\ &\quad +\frac{\kappa\|e\|}{(1-\|A^\dagger\|\|E\|)\|A\|\|x^*\|} \\ &\le \frac{\|q\|}{\|x^*\|}+\kappa\left[\frac{\|p\|}{\|b\|}+\frac{1+\sqrt{5}}{2}\frac{\|A^\dagger\|}{1-\|A^\dagger\|\|E\|}\right)\|E\| \\ &\quad +\frac{\|e\|}{\|b\|(1-\|A^\dagger\|\|E\|)}\Bigg]. \qquad \square \end{aligned}$$

Remark 2.1. In the special case that A^{-1} exists, the feasible set of (1) just consists of $\{x^*\}$, and (4) is reduced to

$$\frac{\|y^*-x^*\|}{\|x^*\|} \le \frac{\kappa}{1-\|A^{-1}E\|}\left(\frac{\|e\|}{\|b\|}+\frac{\|E\|}{\|A\|}\right),$$

which is a classic result in numerical linear algebra.

Remark 2.2. It was not endeavored to get an optimal upper bound for (5). As pointed out by one referee, using $p-x^* = A^\dagger(Ap-b)$, from (8) of [4] a bound slightly different and a little better than (5) can be obtained.

3. Error bound: $b \notin R(A)$. Now we drop the assumption that $b \in R(A)$ and $b + e \in R(A + E)$.

THEOREM 3.1. *Suppose* $\|A^\dagger\|\|E\| < 1$. *Then there is a feasible solution* x *to* (1) *such that*

$$\frac{\|y^* - x\|}{\|x\|} \le \frac{\kappa}{1 - \|A^\dagger E\|} \left[\frac{\|A^\dagger\|\|Ax - b\|\|E\| + 2\|e\|}{\|A\|\|x\|} + 2\frac{\|E\|}{\|A\|} \right]. \tag{7}$$

If, in addition, $\mathrm{Rank}(A + E) = \mathrm{Rank}(A)$ *and* $\|A^\dagger\|\|E\| < 1$, *then*

$$\begin{aligned} \frac{\|y^* - x^*\|}{\|x^*\|} \le \frac{\|q\|}{\|x^*\|} + \kappa \Bigg[& \frac{\|p\|}{\|A\|\|x^*\|} + \frac{1 + \sqrt{5}}{2} \frac{\|A^\dagger\|\|b\|}{(1 - \|A^\dagger\|\|E\|)\|A\|\|x^*\|} \Bigg) \|E\| \\ & + \frac{\|e\|}{(1 - \|A^\dagger\|\|E\|)\|A\|\|x^*\|} \Bigg]. \end{aligned} \tag{8}$$

Proof. Let x be the orthogonal projection of y^* onto the feasible set of (1). Denote

$$r = (A + E)y^* - (b + e).$$

Then

$$(A + E)^\dagger r = 0 \tag{9}$$

and

$$\begin{aligned} \|r\| = \|(A + E)y^* - (b + e)\| &\le \|(A + E)x - (b + e)\| \\ &\le \|Ax - b\| + \|e - Ex\|, \end{aligned} \tag{10}$$

since y^* is a least squares solution of the system $(A + E)y = b + e$. Thus,

$$A^\dagger r = \left[A^\dagger - (A + E)^\dagger\right] r, \tag{11}$$

from which, together with

$$y^* - x = A^\dagger(Ay^* - b) = A^\dagger(r + e - Ey^*) = A^\dagger \left[r + e - E(y^* - x) - Ex\right],$$

$$\begin{aligned} (I + A^\dagger E)(y^* - x) &= A^\dagger r + A^\dagger(e - Ex) \\ &= \left[A^\dagger - (A + E)^\dagger\right] r + A^\dagger(e - Ex). \end{aligned} \tag{12}$$

On the other hand, from the decomposition (see Theorem 8.5 of [5])

$$\begin{aligned} A^\dagger - (A + E)^\dagger = {} & A^\dagger E(A + E)^\dagger - A^\dagger(A^\dagger)^T E^T[I - (A + E)(A + E)^\dagger] \\ & - (I - A^\dagger A)E^T[(A + E)^\dagger]^T(A + E)^\dagger \end{aligned}$$

and (9), we obtain

$$\left[A^\dagger - (A + E)^\dagger\right] r = -A^\dagger(A^\dagger)^T E^T r,$$

from which it follows that

$$\|[A^\dagger - (A + E)^\dagger]r\| \le \|A^\dagger\|^2\|E\|\|r\|. \tag{13}$$

Therefore, by (12), (13), and (10),

$$\begin{aligned}
\frac{\|y^* - x\|}{\|x\|} &\le \frac{\|A^\dagger\|}{1 - \|A^\dagger E\|} \frac{\|A^\dagger\|\|E\|\|r\| + \|e - Ex\|}{\|x\|} \\
&\le \frac{\|A^\dagger\|}{1 - \|A^\dagger E\|} \frac{\|A^\dagger\|\|E\|(\|Ax - b\| + \|e - Ex\|) + \|e - Ex\|}{\|x\|} \\
&\le \frac{\|A^\dagger\|}{1 - \|A^\dagger E\|} \frac{\|A^\dagger\|\|E\|(\|Ax - b\| + \|e\| + \|Ex\|) + \|e\| + \|Ex\|}{\|x\|} \\
&\le \frac{\kappa}{1 - \|A^\dagger E\|} \left[\frac{\|A^\dagger\|\|E\|\|Ax - b\| + \|e\|(\|A^\dagger\|\|E\| + 1)}{\|A\|\|x\|} \right. \\
&\quad \left. + \frac{\|E\|(\|A^\dagger\|\|E\| + 1)}{\|A\|} \right] \\
&\le \frac{\kappa}{1 - \|A^\dagger E\|} \left[\frac{\|A^\dagger\|\|Ax - b\|\|E\| + 2\|e\|}{\|A\|\|x\|} + 2\frac{\|E\|}{\|A\|} \right].
\end{aligned}$$

This proves (7), and (8) can be proved in the same way as for (5). ∎

Remark 3.1. Actually we have proved that, for any feasible solution y of (3), there is a feasible solution x of (1) such that

$$\frac{\|y^* - x\|}{\|x\|} \le \frac{\kappa}{1 - \|A^\dagger E\|} \left[\frac{\|A^\dagger\|\|Ax - b\|\|E\| + 2\|e\|}{\|A\|\|x\|} + 2\frac{\|E\|}{\|A\|} \right].$$

Remark 3.2. Another upper bound for the rank-preserving perturbation was given in [4] (see Theorem 3.1 in [4]).

COROLLARY 3.1. *If, in addition,* $b \in R(A)$, *then*

$$\frac{\|y^* - x\|}{\|x\|} \le \frac{2\kappa}{1 - \|A^\dagger E\|} \left(\frac{\|e\|}{\|b\|} + \frac{\|E\|}{\|A\|} \right). \tag{14}$$

Remark 3.3. The difference between (14) and (4) is that here the perturbed system of linear equations in the problem (3) may not be consistent, while the perturbation in Theorem 2.1 keeps the system consistent.

REFERENCES

[1] M. ARIOLI, A. LARATTA, AND O. MENCHI, *Numerical computation of the projection of a point onto a polyhedron*, J. Optim. Theory Appl., 43 (1984), pp. 495–525.

[2] M. ARIOLI AND A. LARATTA, *Error analysis of algorithms for computing the projection of a point onto a linear manifold*, Linear Algebra Appl., 82 (1986), pp. 1–26.

[3] S. L. CAMPBELL AND C. D. MEYER, JR., *Generalized Inverses of Linear Transformations*, Pitman, Boston, MA, 1979.

[4] J. DING, *Perturbation analysis for the projection of a point to an affine set*, Linear Algebra Appl., 191 (1993), pp. 199–212.

[5] C. L. LAWSON AND R. J. HANSON, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[6] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[7] M. WEI, *On the error estimate for the projection of a point onto a linear manifold*, Linear Algebra Appl., 133 (1990), pp. 53–75.

[8] Y. YE, *A $O(n^3L)$ potential reduction algorithm for linear programming*, Math. Programming, 50 (1991), pp. 239–258.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 701–719, July 1998

007

ON THE QUALITY OF SPECTRAL SEPARATORS*

STEPHEN GUATTERY† AND GARY L. MILLER‡

Abstract. Computing graph separators is an important step in many graph algorithms. A popular technique for finding separators involves spectral methods. However, there has not been much prior analysis of the quality of the separators produced by this technique; instead it is usually claimed that spectral methods "work well in practice." We present an initial attempt at such an analysis. In particular, we consider two popular spectral separator algorithms and provide counterexamples showing that these algorithms perform poorly on certain graphs. We also consider a generalized definition of spectral methods that allows the use of some specified number of the eigenvectors corresponding to the smallest eigenvalues of the Laplacian matrix of a graph, and we show that if such algorithms use a constant number of eigenvectors, then there are graphs for which they do no better than using only the second smallest eigenvector. Furthermore, using the second smallest eigenvector of these graphs produces partitions that are poor with respect to bounds on the gap between the isoperimetric number and the cut quotient of the spectral separator. Even if a generalized spectral algorithm uses n^ϵ for $0 < \epsilon < \frac{1}{4}$ eigenvectors, there exist graphs for which the algorithm fails to find a separator with a cut quotient within $n^{\frac{1}{4}-\epsilon} - 1$ of the isoperimetric number. We also introduce some facts about the structure of eigenvectors of certain types of Laplacian and symmetric matrices; these facts provide the basis for the analysis of the counterexamples. Finally, we discuss some developments in spectral partitioning that have occurred since these results first appeared.

Key words. graph partitioning, spectral partitioning, graph eigenvalues and eigenvectors

AMS subject classifications. 05C50, 05C85, 15A18, 68Q25, 68R10

PII. S0895479896312262

1. Introduction.

Spectral methods (i.e., methods that use the eigenvalues and eigenvectors of a matrix representation of a graph) are widely used to compute graph separators. Typically, the Laplacian matrix is used; the Laplacian B of a graph G on n vertices is the $n \times n$ matrix with the degrees of the vertices of G on the diagonal and entry $b_{ij} = -1$ if G has the edge (v_i, v_j) and 0 otherwise. The eigenvector $\mathbf{u}_2$ corresponding to λ_2 (the second smallest eigenvalue of B) is computed, and the vertices of the graph are partitioned according to the values of their corresponding entries in $\mathbf{u}_2$ [24, 18]. The goal is to compute a small separator; that is, as few edges or vertices as possible should be deleted from the graph to achieve the partition. Additionally, the sizes of the resulting components should be roughly comparable.

Although spectral methods are popular, there has been little previous analysis of the quality of the separators they produce. Instead, it is often claimed that such methods "work well in practice" and tables of results for specific examples are often included in papers (see, e.g., [24]). Thus someone wishing to compute separators has little guidance in determining if this technique is appropriate. Ideally, practitioners should have a characterization of classes of graphs for which spectral separator techniques work well; this characterization might be in terms of how far the computed

*Received by the editors November 3, 1996; accepted for publication (in revised form) by J. W. H. Liu May 17, 1997; published electronically March 18, 1998. An extended abstract of this paper appeared in Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '95), SIAM, Philadelphia, PA, 1995, pp. 233–242. This paper is a revised version of CMU Computer Science Technical Report CMU-CS-94-228, Carnegie Mellon University, Pittsburgh, PA.

http://www.siam.org/journals/simax/19-3/31226.html

†ICASE, Mail Stop 403, NASA Langley Research Center, Hampton, VA 23681 (smg@icase.edu). The research of this author was partially supported by NAS contract NAS1-19480.

‡School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 (glmiller@cs.cmu.edu). The research of this author was partially supported by NSF grant CCR-9505472.

separators can be from optimal. This paper represents a first step in this direction. We consider two spectral separation algorithms that partition the vertices on the basis of the values of their corresponding entries in $\mathbf{u}_2$, and we provide counterexamples from classes of practical interest for which each of the algorithms produces poor separators. We further consider a generalized definition of spectral methods that allows the use of more than one of the eigenvectors corresponding to the smallest nonzero eigenvalues, and we show that there are graphs for which any such algorithm does poorly.

The first algorithm bisects a graph by partitioning the vertices into two equal-sized sets based on each vertex's entry in the eigenvector $\mathbf{u}_2$. The class of bounded-degree planar counterexamples for this method consists of graphs that look like ladders with the top half of their rungs removed; a straightforward spectral bisection algorithm cuts the remaining rungs, whereas the optimal bisection is made by cutting across the ladder above the remaining rungs. The counterexample graphs have $\Theta(1)$ bisectors; the spectral bisection algorithm produces a $\Theta(n)$ bisection, which is as far from the optimum as possible (to within a constant).

The spectral bisection algorithm can be modified to generate a better separator for the bisection counterexample. Some modifications are presented in [18]; they still use a partition based on $\mathbf{u}_2$. We consider a simple spectral separator algorithm, the "best threshold cut" algorithm, based on the most general of these suggested modifications. (In such an algorithm, "best" is measured in terms of the **cut quotient**, the ratio between the number of edges cut and the size of the smaller set in the vertex partition; the smallest cut quotient over all separators is called the **isoperimetric number**.) We present a class of graphs that defeats this algorithm in that the ratio of the spectral cut's cut quotient to the isoperimetric number is as bad as possible (to within a constant) with respect to bounds on these quantities.

We also consider a more general definition of purely spectral separator algorithms that subsumes the two preceding algorithms. This definition allows the use of some specified number of eigenvectors corresponding to the smallest eigenvalues of the Laplacian. For any such algorithm that uses a fixed number of eigenvectors we show there are graphs for which it does no better than using the "best threshold cut" algorithm. Furthermore, the separator produced when the "best threshold cut" algorithm is applied to these graphs is as bad as possible (to within a constant) with respect to bounds on the size of the separators produced. We also show that if a purely spectral algorithm uses up to n^ϵ eigenvectors for $0 < \epsilon < \frac{1}{4}$, there exist graphs for which the algorithm fails to find a separator with a cut quotient within a factor of $n^{\frac{1}{4}-\epsilon} - 1$ times the isoperimetric number.

Finally, we provide a summary of some important subsequent results by Spielman and Teng [28] and relate our results to them.

This paper makes an additional contribution. While the counterexamples have simple structures and intuitively might be expected to cause problems for spectral separator algorithms, the challenge is to provide good bounds on λ_2 for these graphs. For this purpose we have developed theorems about the spectra of graphs with particular symmetries (i.e., automorphisms of order 2) that exist in the counterexamples.

Specifics are given in the text that follows. Section 2 gives a brief history of spectral methods and the details of the algorithms discussed in this paper. Graph and matrix terminology and notation are presented in section 3, which also presents some useful facts about Laplacians. Results about the eigenvalues and eigenvectors of Laplacians of graphs with automorphisms of order 2 are in section 4. Section 5 gives the counterexample for the spectral bisection algorithm; section 6 gives the counterex-

ample for the "best threshold cut" algorithm. Section 7 discusses the generalized definition of spectral separator algorithms and shows that there are graphs for which any such algorithm performs poorly. Section 8 discusses the results of Spielman and Teng.

2. Spectral methods for computing separators. The roots of spectral partitioning go back to Donath and Hoffman [9], who proved a lower bound on the size of the minimum bisection of a graph, and Fiedler [11][12], who explored the properties of λ_2 and its associated eigenvector for the Laplacian. There has been much subsequent work, including Barnes's partitioning algorithm [5], Boppana's work that included a stronger lower bound on the minimum bisection size [6], work by Rendl, Wolkowicz, and others using optimization approaches [25], [10], and the particular bisection and graph partitioning methods considered in this paper [18], [24], [26]. Since our work first appeared [17], Spielman and Teng [28] have extended the latter methods to include recursion. (It is worth noting that spectral methods have not been limited to graph partitioning; work has been done using the spectrum of the adjacency matrix in graph coloring [4] and using the Laplacian spectrum to prove theorems about expander graph and superconcentrator properties [3], [1], [2]. The work on expanders has explored the relationship of λ_2 to the isoperimetric number; Mohar has given an upper bound on the isoperimetric number using a strong discrete version of the Cheeger inequality [23]. Reference [8] is a book-length treatment of graph spectra, and it predates many of the results cited above.)

A basic way of computing a graph bisection using spectral information is presented in [24]. We refer to this algorithm as **spectral bisection**. Spectral bisection works as follows:

- Represent G by its Laplacian B, and compute $\mathbf{u}_2$, the eigenvector corresponding to λ_2 of B.
- Assign each vertex the value of its corresponding entry in $\mathbf{u}_2$. This is the **characteristic valuation** of G.
- Compute the median of the elements of $\mathbf{u}_2$. Partition the vertices of G as follows: the vertices whose values are less than or equal to the median form one part; the rest of the vertices form the other part. The set of edges between the two parts forms an edge separator.
- If a vertex separator is desired, it is computed from the edge separator using standard techniques described in the next section.

Since the graph bisection problem is NP-complete [13], spectral bisection may not give an optimum result. That is, spectral bisection is a heuristic method. A number of modifications have been proposed that may improve its performance. These modified heuristics may give splits other than bisections. In such cases, one can use the cut quotient to judge the quality of the split. Computing a separator with a cut quotient equal to the isoperimetric number is NP-hard [14]. The following modifications, all of which use the characteristic valuation, are presented in [18]:

- Partition the vertices based on the signs of their values;
- look for a large gap in the sorted list of eigenvector components, and partition the vertices according to whether their values are above or below the gap; and
- sort the vertices according to value. For each index $1 \leq i \leq n-1$, consider the ratio for the separator produced by splitting the vertices into those with sorted index $\leq i$ and those with sorted index $> i$. Choose the split that provides the best cut quotient.

Note that the last idea subsumes the first two. We consider a variant of this algorithm below. Since this algorithm does not specify what to do when multiple vertices have

the same value, we restrict it to consider only splits between vertices with different values (such cuts are called **threshold cuts**). This restricted version is the **"best threshold cut" algorithm**; the slight change from the definition above does not alter its performance with respect to the counterexamples below (other than slightly simplifying the analysis).

Also note that the idea of cutting at an arbitrary point along the sorted order can be extended to choosing two split points, where the corresponding partitions are the vertices with values between the split points, and those with values above the upper or below the lower split point. Again, the pair yielding the best ratio is chosen.

The algorithms mentioned so far have used only the eigenvector $\mathbf{u}_2$. Another possibility is to look at partitions generated by the set of eigenvectors for some number of smallest eigenvalues: for each vertex, a value is assigned by computing a function of that vertex's eigenvector components. Partitions are then generated in the same way as they are for $\mathbf{u}_2$ in the various algorithms given above.

Given the variety of heuristics cited above, it would be nice to know which ones work well for which classes of graphs. It would be particularly useful if it were possible to state reasonable bounds on the performance of these heuristics for classes of graphs commonly used in practice (e.g., planar graphs, planar graphs of bounded degree, three-dimensional finite element meshes, etc.). Unfortunately, this is not the case. We start by proving in section 5 that spectral bisection may produce a bad separator for a bounded-degree planar graph; first, however, we need to introduce some terminology and background results.

3. Terminology, notation, and background results. We assume that the reader is familiar with the basic definitions of graph theory (in particular, for undirected graphs) and with the basic definitions and results of matrix theory. A graph consists of a set of vertices V and a set of edges E; we denote the vertices (respectively, edges) of a particular graph G as $V(G)$ (respectively, $E(G)$) if there is any ambiguity about which graph is being referred to. The notation $|G|$ is sometimes used as a shorthand for $|V(G)|$. When it is clear which graph is being referred to, we use n to denote $|V|$.

Capital letters represent matrices and bold lowercase letters represent vectors. For a matrix A, a_{ij} or $[A]_{ij}$ represents the element in row i and column j; for the vector $\mathbf{x}$, x_i or $[\mathbf{x}]_i$ represents the ith entry in the vector. The notation $\mathbf{x} = 0$ indicates that all entries of the vector $\mathbf{x}$ are zero; $\vec{1}$ indicates the vector that has 1 for every entry. For ease of reference, the eigenvalues of an $n \times n$ matrix are indexed in nondecreasing order. λ_1 represents the smallest eigenvalue, and λ_n represents the largest. For $1 < i < n$, $\lambda_{i-1} \leq \lambda_i \leq \lambda_{i+1}$. The notation $\lambda_i(A)$ (respectively, $\lambda_i(G)$) indicates the ith eigenvalue of matrix A (respectively, of the Laplacian of graph G) if there is any ambiguity about which matrix (respectively, graph) the eigenvalue belongs to. We use $\mathbf{u}_i$ to represent the eigenvector corresponding to λ_i.

A **path graph** is a tree with exactly two vertices of degree one.

The **crossproduct** of two graphs G and H (denoted $G \times H$) is a graph on the vertex set $\{(u, v) \mid u \in V(G), v \in V(H)\}$, with $((u, v), (u', v'))$ in the edge set if and only if either $u = u'$ and $(v, v') \in E(H)$ or $v = v'$ and $(u, u') \in E(G)$. It is easy to see that $G \times H$ and $H \times G$ are isomorphic. One way to think of a graph crossproduct is as follows. Replace every vertex in G with a copy of H. Each edge e in G is then replaced by $|H|$ edges, one between each pair of corresponding vertices in the copies of H that have replaced the endpoints of e. An example is shown in Figure 3.1.

For a connected graph G, an **edge separator** is a set S of edges that, if removed,

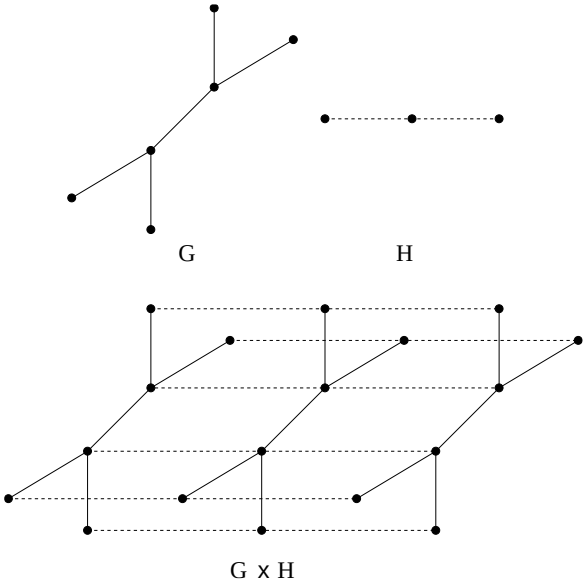


FIG. 3.1. *A graph crossproduct example.*

breaks the graph into two (not necessarily connected) components G_1 and G_2 that have no edges between them. (An edge separator is by definition minimal with respect to G_1 and G_2.) A **vertex separator** is a set S of vertices such that if these vertices and all incident edges are removed, the graph is broken into two components G_1 and G_2 that have no edges between them (again, a separator is a minimal such set). The goal in computing separators is to find a small set S that breaks the graph into two fairly large pieces; often this notion is expressed as a balance restriction that requires the number of vertices in each of G_1 and G_2 to be at least some specified fraction of the number of vertices in G. For edge separators, this goal is stated more generally in terms of minimizing some measure relating the size of the separator to the size of the resulting components. One such measure that we use is the **isoperimetric number** $i(G)$, defined as

$$\min_S \left(\frac{|S|}{\min(|G_1|, |G_2|)} \right).$$

We refer to the quantity $|S|/\min(|G_1|, |G_2|)$ as the **cut quotient** for the edge separator S. As noted in section 2, finding a cut with a cut quotient equal to the isoperimetric number is NP-hard. It is well known that an edge separator S can be converted into a vertex separator S' by considering the bipartite graph induced by S and setting S' to be a minimum vertex cover for that graph.

Given a vertex numbering, graphs can be represented by matrices. For example, the **adjacency matrix** A of a graph G is defined as $a_{ij} = 1$ if and only if $(v_i, v_j) \in E(G)$; $a_{ij} = 0$ otherwise. A common matrix representation of graphs is the **Laplacian**. Let D be the matrix with $d_{ii} = \text{degree}(v_i)$ for $v_i \in V(G)$, and all off-diagonal entries equal to zero. Let A be the adjacency matrix for G. Then the Laplacian of G is the matrix $B = D - A$.

The following are useful facts about the Laplacian matrix:

- The Laplacian is symmetric positive semidefinite (see, e.g., [22]).

- A graph G is connected if and only if 0 is a simple eigenvalue of its Laplacian (see, e.g., [22]). The eigenvector for 0 is $\vec{1}$.
- The following characterization of λ_2 holds (see, e.g., [11]):

$$\lambda_2 = \min_{\mathbf{x} \perp \vec{1}} \frac{\mathbf{x}^T B \mathbf{x}}{\mathbf{x}^T \mathbf{x}}.$$

- If G is a crossproduct of two graphs H_1 and H_2, then the eigenvalues of the Laplacian of G are all pairwise sums of the eigenvalues of H_1 and H_2 (see, e.g., [22]).
- For any vector $\mathbf{x}$ and Laplacian B of graph G, the following holds (see, e.g., [18]):

$$\mathbf{x}^T B \mathbf{x} = \sum_{(v_i, v_j) \in E(G)} (x_i - x_j)^2. \tag{3.1}$$

- For a graph G that is not one of K_1, K_2, or K_3 (the complete graphs on 1, 2, and 3 vertices, respectively), let λ_2 be the smallest nonzero eigenvalue of its Laplacian. G's isoperimetric number can be bounded as follows [23]:

$$\frac{\lambda_2}{2} \leq i(G) \leq \sqrt{\lambda_2(2\Delta - \lambda_2)}, \tag{3.2}$$

  where Δ is the maximum degree of any vertex in G.

The proof of the upper bound in (3.2) has interesting implications about the threshold cuts based on the second eigenvector. For any connected graph G, consider the characteristic valuation. The vertices of G receive $k \leq n$ distinct values; let $t_1 > t_2 > \cdots > t_k$ be these values. For each threshold t_i, $i < k$, divide the vertices into those with values greater than t_i and those with values less than or equal to t_i. Compute the cut quotient q_i for each such cut, and let q_{min} be the minimum over all q_i's. The following theorem can be derived from the proof of Theorem 4.2 in [23] (a similar argument leading to similar result for the Laplace operator associated with the transition matrix of a reversible Markov chain can be found in [27]).

THEOREM 3.1. *Let G be a connected graph with maximal vertex degree Δ and second smallest eigenvalue λ_2. If G is not any of K_1, K_2, or K_3, then*

$$\frac{\lambda_2}{2} \leq q_{min} \leq \sqrt{\lambda_2(2\Delta - \lambda_2)}.$$

A **weighted** graph is a graph for which a real value w_i is associated with each vertex v_i, and a real, nonzero weight w_{ij} is associated with each edge (v_i, v_j) (a zero edge weight indicates the lack of an edge). Fiedler extended the notion of the Laplacian to graphs with positive edge weights [12]; he referred to this representation as the **generalized Laplacian**. Our results require a representation for graphs with vertex weights and negative edge weights. Hence we define the **standard matrix representation** B of a weighted graph G as follows: B has $b_{ii} = w_i$; for $i \neq j$ and $(v_i, v_j) \in E(G)$, $b_{ij} = -w_{ij}$, and $b_{ij} = 0$ otherwise. Note that the standard matrix representation of any weighted graph is a real symmetric matrix and that any such matrix can be represented as a specific weighted graph. Note also that the Laplacian matrix of a graph is also the standard matrix representation of the graph with vertex weights equal to the vertex degrees and all edge weights set to 1.

4. Automorphisms of order 2 and eigenvector structure. The theorems and lemmas presented in this section are useful in proving results about the eigenvectors of the families of graphs presented in later sections. The details of the proofs are not necessary to understand the rest of the paper; a reader interested only in understanding the counterexamples and their implications can look at the theorem statements and skip the proofs.

The first set of results concerns eigenvalues of Laplacians of graphs with automorphisms of order 2. A **graph automorphism** is a permutation ϕ on the vertices of the graph G such that $(v_i, v_j) \in E(G)$ if and only if $(v_{\phi(i)}, v_{\phi(j)}) \in E(G)$. The **order** of a graph automorphism is the order of the permutation ϕ, the minimum number of times ϕ must be applied to yield the identity mapping.

For weighted graphs, there are two additional conditions: the weights of vertices v_i and $v_{\phi(i)}$ must be equal for all i, and the weights of edges (v_i, v_j) and $(v_{\phi(i)}, v_{\phi(j)})$ must be equal.

The next two theorems concern the structure of eigenvectors with respect to automorphisms of order 2. They hold for both Laplacians of graphs under the standard definition of automorphism and standard matrix representations of weighted graphs under the definition of automorphisms for weighted graphs.

Let G be a graph with an automorphism ϕ of order 2 and Laplacian B. A vector $\mathbf{x}$ that has $x_i = x_{\phi(i)}$ for all i in the range $1 \leq i \leq n$ is an **even** vector with respect to the automorphism ϕ; an **odd** vector $\mathbf{y}$ has $y_i = -y_{\phi(i)}$ for all i. It is easy to show that for any even vector $\mathbf{x}$ and odd vector $\mathbf{y}$ (both with respect to ϕ), $\mathbf{x}$ and $\mathbf{y}$ are orthogonal.

THEOREM 4.1. *Let B be the Laplacian of a graph G that has an automorphism ϕ of order 2. Then there exists a complete set $\mathcal{U}$ of orthogonal eigenvectors of B such that any eigenvector $\mathbf{u} \in \mathcal{U}$ is either even or odd with respect to ϕ. This also holds if G is a weighted graph, B is the standard matrix representation of G, and ϕ is a weighted graph automorphism of order 2.*

Proof. Let P be the permutation matrix that corresponds to the automorphism ϕ. Then $P^T B P = B$. Let $\mathbf{u}$ be an eigenvector of B with eigenvalue λ. We have

$$\left(P^T B P\right)\mathbf{u} = B\mathbf{u} = \lambda\mathbf{u}. \tag{4.1}$$

Since the automorphism is of order 2, $PP = I$ and $P^T = P^{-1} = P$. Therefore, multiplying the left and right sides of (4.1) by P gives

$$B\left(P\mathbf{u}\right) = P\left(\lambda\mathbf{u}\right) = \lambda\left(P\mathbf{u}\right).$$

Thus, $P\mathbf{u}$ is also an eigenvector with eigenvalue λ.

Note that for an even vector $\mathbf{x}$, $P\mathbf{x} = \mathbf{x}$; for an odd vector $\mathbf{y}$, $P\mathbf{y} = -\mathbf{y}$.

P allows us to decompose any vector $\mathbf{x}$ uniquely into an odd component $\mathbf{x}_{odd}$ and an even component $\mathbf{x}_{even}$ as follows:

$$\mathbf{x}_{odd} = \frac{\mathbf{x} - P\mathbf{x}}{2}, \quad \text{and} \quad \mathbf{x}_{even} = \frac{\mathbf{x} + P\mathbf{x}}{2}.$$

For any nonzero $\mathbf{x}$, at least one of the even or odd parts must also be nonzero.

Let $\mathcal{U}'$ be any complete set of eigenvectors of B. For an eigenvector $\mathbf{u} \in \mathcal{U}'$, it is easy to see that a nonzero even or odd component is an eigenvector for the same eigenvalue. Since $\mathbf{u}_{odd} + \mathbf{u}_{even} = \mathbf{u}$, the set of odd and even eigenvectors resulting from decomposing all the eigenvectors in $\mathcal{U}'$ spans the same space as $\mathcal{U}'$.

The subspaces spanned by all odd and by all even components, respectively, are orthogonal. Since B is real and symmetric, we can subdivide these subspaces into smaller orthogonal subspaces spanned by the odd (respectively, even) eigenvectors for particular eigenvalues. We can form an orthogonal basis for each of these smaller subspaces; the union of all these bases is the desired set $\mathcal{U}$ of orthogonal odd and even eigenvectors. This implies the claimed result.

The proof clearly holds whether B is a Laplacian or a standard matrix representation. □

COROLLARY 4.2. *Let B be the standard matrix representation of a weighted graph G that has one or more automorphisms of order 2. Then the eigenvector for any simple eigenvalue is either even or odd with respect to every such automorphism.*

Proof. Let $\mathbf{u}$ be the eigenvector for some simple eigenvalue λ. Consider the decomposition of $\mathbf{u}$ into odd and even parts with respect to some automorphism ϕ with order 2. If both parts were nonzero, they would be orthogonal and eigenvectors for λ. Therefore either the odd part or the even part must be zero. □

Since Laplacians can be considered standard matrix representations given the right weight assignments, the preceding result also holds for Laplacians.

Let B be a standard matrix representation of a weighted graph with an automorphism ϕ of order 2. It is possible to decompose B into two smaller matrices B_{odd} and B_{even} such that the eigenvalues of B_{odd} and B_{even} are the odd and even eigenvalues of B, respectively, and furthermore that a full set of odd and even eigenvectors of B can be constructed in a simple way from the eigenvectors of B_{odd} and B_{even}, respectively. We demonstrate this through a similarity transform based on ϕ. First, however, we need to introduce some notation.

The vertices of G can be divided into two disjoint sets on the basis of how ϕ operates on them. Let V_f be the set of vertices v_i such that $\phi(i) = i$ (i.e., the vertices fixed by ϕ) and let V_m be the set of vertices v_j such that $\phi(j) \neq j$ (i.e., the vertices moved by ϕ). V_m consists of vertices in orbits of length 2. We call a subset of V_m that consists of exactly one vertex from each such orbit a **representative set** and denote it V_r. In the rest of this presentation we assume that a particular V_r has been arbitrarily specified. We use n_f, n_m, and n_r, respectively to denote the number of vertices in each of these sets.

Without loss of generality, number the vertices in the following way: the vertices in V_f are numbered 1 through n_f; the vertices in V_r are numbered from $n_f + 1$ to $n_f + n_r$. Renumber the vertices in $V_m \setminus V_r$ such that if $v_i \in V_r$, then $\phi(i) = i + n_r$; that is, the vertices in $V_m \setminus V_r$ are numbered $n_f + n_r + 1$ to n in the same order as the vertices in V_r with which they share orbits. Using this ordering and the definition of the automorphism, B can be written in the following block form:

$$B = \begin{bmatrix} F & E_{fr} & E_{fr} \\ E_{fr}^T & R & E_{r\phi(r)} \\ E_{fr}^T & E_{r\phi(r)} & R \end{bmatrix},$$

where

- F is an $n_f \times n_f$ submatrix containing the diagonal entries for the vertices in V_f and the entries for edges between pairs of vertices in V_f;
- R is an $n_r \times n_r$ submatrix containing the diagonal entries for the vertices in V_r and the entries for edges between pairs of vertices in V_r;
- E_{fr} is made up of the entries of B for edges between vertices in V_f and V_r; and

- $E_{r\phi(r)}$ is made up of the entries of B for edges between vertices in V_r and $V_f \setminus V_r$ (note that the conditions specified above imply $E_{r\phi(r)} = E_{r\phi(r)}^T$).

We now define the orthogonal matrix T used to transform B:

$$T = \begin{bmatrix} I_{n_f} & 0 & 0 \\ 0 & \frac{1}{\sqrt{(2)}} I_{n_r} & \frac{1}{\sqrt{(2)}} I_{n_r} \\ 0 & \frac{1}{\sqrt{(2)}} I_{n_r} & \frac{-1}{\sqrt{(2)}} I_{n_r} \end{bmatrix},$$

where the I's are identity matrices with the dimension specified in the subscript. B is transformed as follows:

$$B' = T^T B T = \begin{bmatrix} F & \sqrt{2}E_{fr} & 0 \\ \sqrt{2}E_{fr}^T & R + E_{r\phi(r)} & 0 \\ 0 & 0 & R - E_{r\phi(r)} \end{bmatrix}.$$

Note that the resulting matrix is reducible. That is, when viewed as a weighted graph, that graph has two components. We show that the blocks of this matrix correspond to B_{even} and B_{odd} as follows:

$$B_{even} = \begin{bmatrix} F & \sqrt{2}E_{fr} \\ \sqrt{2}E_{fr}^T & R + E_{r\phi(r)} \end{bmatrix} \quad \text{and} \quad B_{odd} = R - E_{r\phi(r)}.$$

Let B, T, B', B_{odd}, and B_{even} be as defined above.

THEOREM 4.3. *The eigenvalues of B_{odd} are odd eigenvalues of B, and a complete set of odd eigenvectors of B can be constructed from the eigenvectors of B_{odd} in a straightforward way. Likewise, the eigenvalues of B_{even} are even eigenvalues of B, and a complete set of even eigenvectors of B can be constructed from the eigenvectors of B_{even} in a straightforward way.*

Proof. Because B' is reducible, every eigenvalue of B_{odd} is an eigenvalue of B'; likewise every eigenvalue of B_{even} is an eigenvalue of B'. By similarity, they are also eigenvalues of B.

Now consider an eigenvector $\mathbf{u}$ of B_{even}. Define $\mathbf{v}$ as follows: for $1 \le i \le n_f + n_r$ let $v_i = u_i$; let $v_i = 0$ otherwise. $\mathbf{v}$ is obviously an eigenvector of B'. Multiplication by the matrix T transforms $\mathbf{v}$ into an eigenvector $\mathbf{w}$ of B as follows:

$$\mathbf{w} = T\mathbf{v} = \begin{bmatrix} \mathbf{v}_f \\ \frac{1}{\sqrt{(2)}} \mathbf{v}_r \\ \frac{1}{\sqrt{(2)}} \mathbf{v}_r \end{bmatrix}.$$

By the vertex numbering, it is easy to see that this is an even vector. Since $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ all have the same eigenvalue λ, the claim about eigenvalues of B_{even} corresponding to even eigenvalues of B holds. It is easy to show that if two eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ of B_{even} are orthogonal, then the corresponding eigenvectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are also orthogonal. Since B_{even} has $n_f + n_r$ orthogonal eigenvectors, we have $n_f + n_r$ orthogonal even eigenvectors of B.

Now consider an eigenvector $\mathbf{u}$ of B_{odd}. As before, one can construct an eigenvector $\mathbf{v}$ of B': for $n_f + n_r + 1 \le i \le n$ let $v_i = u_i$; let $v_i = 0$ otherwise. Multiplication by the matrix T again transforms $\mathbf{v}$ into an eigenvector $\mathbf{w}$ of B as follows:

$$\mathbf{w} = T\mathbf{v} = \begin{bmatrix} 0 \\ \frac{1}{\sqrt{(2)}} \mathbf{v}_{\phi(r)} \\ \frac{-1}{\sqrt{(2)}} \mathbf{v}_{\phi(r)} \end{bmatrix}.$$

This is clearly an odd vector. Since $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$ all have the same eigenvalue λ, the claim about eigenvalues of B_{odd} corresponding to odd eigenvalues of B holds. It is easy to show that if two eigenvectors $\mathbf{u}_1$ and $\mathbf{u}_2$ of B_{odd} are orthogonal, then the corresponding eigenvectors $\mathbf{w}_1$ and $\mathbf{w}_2$ are also orthogonal. Since B_{odd} has n_r orthogonal eigenvectors, we have n_r orthogonal odd eigenvectors of B.

Note that if all eigenvectors of B_{even} and B_{odd} are transformed in this way the result is n orthogonal eigenvectors of B (i.e., a full set). □

It is possible to express this decomposition in terms of graphs. The graph G is decomposed into the components G_{odd} and G_{even}. Rules for the graphical decomposition can be derived from the structure of B_{odd} and B_{even} and are presented in the technical report version of this paper [16].

The following technical lemmas about the eigenvalues and eigenvectors of weighted path graphs are useful in subsequent results.

LEMMA 4.4. *Let B be the standard matrix representation of a weighted path graph G on n vertices. For any vector $\mathbf{x}$ such that $B\mathbf{x} = \lambda\mathbf{x}$ for some real λ, $x_n = 0$ implies $\mathbf{x} = 0$. Likewise, $x_1 = 0$ implies $\mathbf{x} = 0$. If there are two consecutive elements x_i and x_{i+1} that are both zero, then $\mathbf{x} = 0$.*

Proof. The first result is proved by induction. The base case is for a 2×2 matrix with diagonal entries b_{11} and b_{22}, and off-diagonal entries $b_{12} = b_{21} = -c$. Let $\mathbf{x}$ and λ be as specified by the lemma statement, and assume that $x_2 = 0$. The second element of the vector resulting from multiplying $B\mathbf{x} = \lambda\mathbf{x}$ is $-c \cdot x_1 = \lambda x_2 = 0$. Since $c \neq 0$ by definition (G is a weighted path graph), it must be the case $x_1 = 0$, which implies that $\mathbf{x} = 0$.

For the induction step, assume that the result holds for all $i \leq k$, and consider the standard matrix representation of a weighted path graph on $k + 1$ vertices. Let the weight of edge (v_k, v_{k+1}) be c. Let $\mathbf{x}$ and λ be as stated, and assume that $x_{k+1} = 0$. Then $[B\mathbf{x}]_{k+1} = -c \cdot x_k = \lambda x_{k+1} = 0$. Thus $x_k = 0$. Let $\mathbf{x}'$ be the subvector of $\mathbf{x}$ consisting of the first k entries. Note that with $x_{k+1} = 0$ it is the case that $\mathbf{x}'$ and λ meet the lemma conditions for the principal leading minor B_k of B, and that $x'_k = 0$. However, B_k is the standard matrix representation for the weighted path graph derived from G by deleting the last edge and vertex. Thus, by the induction hypothesis $\mathbf{x}'$ must be 0; because $x_{k+1} = 0$, this implies that $\mathbf{x} = 0$.

A symmetric argument implies the result for $x_1 = 0$.

Again, let B be the standard matrix representation of a weighted path graph G. Let $\mathbf{x}$ be a vector meeting the lemma conditions for λ, and assume that $\mathbf{x}$ has two consecutive zero elements x_i and x_{i+1}. If either $i = 1$ or $i + 1 = n$, $\mathbf{x} = 0$ by the previous argument. Otherwise, $x_{i+1} = 0$ implies that the first i elements of $\mathbf{x}$ and λ meet the lemma conditions for the leading principal minor B_i of B. Note that B_i is the standard matrix representation for some weighted path graph. Thus by the previous result the first i entries of $\mathbf{x}$ are zero. By a symmetric argument for the trailing principal minor, the last $n - i$ entries must also be zero, which gives $\mathbf{x} = 0$. □

This lemma implies that for eigenvectors of the standard matrix representation of any weighted path graph, neither the first nor the last entry is zero. Likewise, such an eigenvector cannot have two consecutive zero entries. These facts can be used to give a simple proof of the following lemma (for a different proof, see, e.g., pp. 910–911 of [29]).

LEMMA 4.5. *All eigenvalues of the standard matrix representation B of a weighted path graph G on n vertices are simple (i.e., have multiplicity one).*

Proof. Let $\mathbf{u}$ and $\mathbf{u}'$ be any two eigenvectors of B for the eigenvalue λ. By

Lemma 4.4, $u_n \neq 0$ and $u'_n \neq 0$. Let α be u'_n/u_n; α is non-zero and real. Then $B(\alpha\mathbf{u} - \mathbf{u}') = \lambda(\alpha\mathbf{u} - \mathbf{u}')$. But the n^{th} element of $(\alpha\mathbf{u} - \mathbf{u}')$ is 0, so by Lemma 4.4, it must be the case that $\alpha\mathbf{u} = \mathbf{u}'$, so $\mathbf{u}$ must be a scalar multiple of $\mathbf{u}'$; it is not a distinct eigenvector. ∎

A path graph on n vertices has exactly one automorphism of order two: $\phi(i) = n - i + 1$. Thus one can talk about odd and even eigenvectors of a path graph without ambiguity; they are always defined with respect to this automorphism.

LEMMA 4.6. *Let G be an unweighted path graph on n vertices with Laplacian B. The eigenvector $\mathbf{u}_2$ corresponding to $\lambda_2(B)$ is odd.*

Proof. By Lemma 4.5, $\mathbf{u}_2$ is simple, so by Corollary 4.2, $\mathbf{u}_2$ must be either even or odd. Assume that it is even. We show that this leads to a contradiction.

There are two cases to keep track of: n is odd, and n is even. If n is odd, there is a single center vertex $v_{\lceil \frac{n}{2} \rceil}$ (index the vertices along the path from 1 to n). If n is even, there are two center vertices with indices $\frac{n}{2}$ and $\frac{n}{2}+1$; since $\mathbf{u}_2$ is assumed to be even, their entries in $\mathbf{u}_2$ are equal. Thus, by Lemma 4.4, if n is even the eigenvector entries corresponding to the center vertices are nonzero. If n is odd, $\mathbf{u}_2$ is even, and the eigenvector entry for the center vertex is 0, then it is easy to check that changing the signs of all eigenvector entries with index less than the center index gives an odd eigenvector with eigenvalue λ_2, which contradicts the simplicity of λ_2. Thus, the assumption that $\mathbf{u}_2$ is even implies that the eigenvector entries corresponding to the center vertex or vertices must be nonzero. Let this value be c.

Now consider the vector $\mathbf{x} = (-c) \cdot \vec{1} + \mathbf{u}_2$. Recall that $\mathbf{u}_2$ is orthogonal to $\vec{1}$. It is easy to see that $\mathbf{x}$ is even, and since $c \neq 0$,

$$\frac{\mathbf{x}^T B \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \frac{\mathbf{u}_2^T B \mathbf{u}_2}{c^2 n + \mathbf{u}_2^T \mathbf{u}_2} < \frac{\mathbf{u}_2^T B \mathbf{u}_2}{\mathbf{u}_2^T \mathbf{u}_2} = \lambda_2.$$

However, the entries of $\mathbf{x}$ corresponding to the center vertex or vertices are 0, so as above, one can create an odd vector $\mathbf{y}$ such that $\frac{\mathbf{y}^T B \mathbf{y}}{\mathbf{y}^T \mathbf{y}} = \frac{\mathbf{x}^T B \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$ as follows: set $y_i = x_i$, $i < \frac{n}{2}$, and $y_i = -x_i$; $i > \frac{n}{2}$. Recall the characterization $\lambda_2 = \min_{\mathbf{x} \perp \vec{1}} \frac{\mathbf{x}^T B \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$; $\mathbf{y}$ is orthogonal to $\vec{1}$, so it meets the criteria for the characterization of λ_2, so the assumption that $\mathbf{u}_2$ is even gives $\lambda_2 < \lambda_2$, which is a contradiction. ∎

The reader can easily verify that this theorem also holds for generalized Laplacians (i.e., Fiedler's matrix representation of graphs with positive edge weights [12]) where the automorphism ϕ exists. However, extension to the standard matrix representation case is not possible because of vertex weights and negative edge weights.

5. A bad family of bounded-degree planar graphs for spectral bisection. In this section we present a family of bounded-degree planar graphs that have constant-size separators. However, the separators produced by spectral bisection have size $\Theta(n)$ for both edge and vertex separators.

The family of graphs is parameterized on the positive integers. G_k consists of two path graphs, each on $2k$ vertices, with a set of edges between the two paths as follows: label the vertices of one path from 1 to $2k$ in order (the **upper path**), and label the other path from $2k+1$ to $4k$ in order (the **lower path**). For $1 \leq i \leq k$ there is an edge between vertices $k+i$ and $3k+i$. An example for $k = 5$ is shown in Figure 5.1. It is obvious that G_k is planar for any k and that the maximum degree of any vertex is 3.

Note that the graph has the approximate shape of a cockroach, with the section containing edges between the upper and lower paths being the body and the other

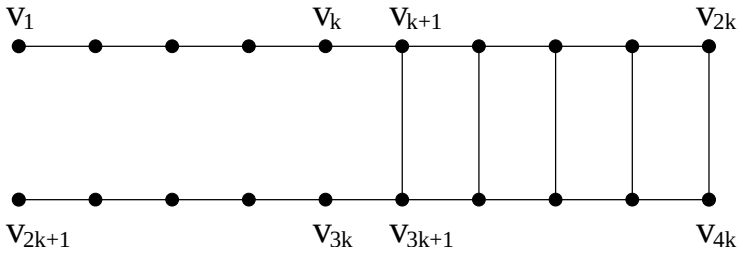


FIG. 5.1. *The roach graph for* $k = 5$.

sections of the paths being antennae. This terminology allows easy references to parts of the graph.

G_k has one automorphism of order 2 that maps the vertices of the upper path to the vertices of the lower path and vice versa. For the rest of this section, the terms "odd vector" and "even vector" are used with respect to this automorphism. Thus, an even vector $\mathbf{x}$ has $x_i = x_{2k+i}$ for all i in the range $1 \leq i \leq 2k$; an odd vector $\mathbf{y}$ has $y_i = -y_{2k+i}$ for all i, $1 \leq i \leq 2k$.

We can now discuss the structure of the eigenvectors of B_k, the Laplacian of G_k.

LEMMA 5.1. *Any eigenvector* $\mathbf{u}_i$ *with eigenvalue* λ_i *of* B_k *can be expressed as a linear combination of*

- *an even eigenvector of* B_k *in which the values associated with the upper path are the same as for the eigenvector with eigenvalue* λ_i *(if it exists) of a path graph on* $2k$ *vertices, and*
- *an odd eigenvector of* B_k *in which the values associated with the upper path are the same as for the eigenvector with eigenvalue* λ_i *(if it exists) of a weighted graph that consists of a path graph on* $2k$ *vertices for which the vertex weights of* v_{k+1} *through* v_{2k} *have been increased by* 2.

Proof. The fact that we can express any eigenvector of B_k as a sum of odd and even eigenvectors follows by Theorem 4.1, applied with respect to the automorphism of order 2.

The claim about the specific structure of the odd and even eigenvectors of B_k follows from an application of the even-odd decomposition proved in Theorem 4.3, with the odd and even matrix components described in graph form. □

It is obvious that G_k has a bisector of constant size: cut the edges connecting the antennae to the body. The following theorem shows that spectral bisection gives much larger bisectors for the family of graphs G_k.

THEOREM 5.2. *Spectral bisection produces* $\Theta(n)$ *edge and vertex separators for* G_k *for any* k.

Proof. The first step is to show that $\mathbf{u}_2$ is odd. Intuitively, this implies that the spectral method splits the graph into the upper path and the lower path.

Recall that $\lambda_2 = \min_{\mathbf{x} \perp \vec{1}} \frac{\mathbf{x}^T B_k \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$. We construct an odd vector $\mathbf{x}$ such that the quotient $\frac{\mathbf{x}^T B_k \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$ is less than $\frac{\mathbf{y}^T B_k \mathbf{y}}{\mathbf{y}^T \mathbf{y}}$ for any even eigenvector $\mathbf{y}$ orthogonal to $\vec{1}$ ($\vec{1}$ is the smallest even eigenvector). This requires a proof that $\frac{\mathbf{x}^T B_k \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$ is less than the second smallest even eigenvalue. From Lemma 5.1 above, the second smallest even eigenvalue of B_k is the same as the second smallest eigenvalue μ_2 of the Laplacian B of a path graph G on $2k$ vertices; it is well known that $\mu_2 = 4\sin^2(\frac{\pi}{4k})$ (see, for example, [22]).

Let $\mathbf{z}$ be the eigenvector of B corresponding to μ_2. Construct $\mathbf{x}$ as follows:

$$x_i = \begin{cases} z_i & 1 \le i \le k, \\ z_{4k-i+1} & 2k+1 \le i \le 3k, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

That is, assign the first k values from the path G to the upper antenna of the roach, working in the direction toward the body and assign the last k entries from G to the lower antenna, working from the body outward. Since $\mathbf{z}$ and $\mathbf{x}$ have the same set of nonzero entries, $\mathbf{x}^T\mathbf{x} = \mathbf{z}^T\mathbf{z}$. Likewise, since $\mathbf{z}$ is orthogonal to the "all-ones" vector, so is $\mathbf{x}$.

To see that $\mathbf{x}^T B_k \mathbf{x} < \mathbf{z}^T B\mathbf{z}$, recall (3.1) from section 3: for Laplacian B and vector $\mathbf{y}$,

$$\mathbf{y}^T B\mathbf{y} = \sum_{(v_i,v_j)\in E} (y_i - y_j)^2.$$

For every edge in G except one, there is an edge in G_k that contributes the same value to this sum. The one exception is the edge (v_k, v_{k+1}) in G. Since $\mathbf{z}$ is an odd vector by Lemma 4.6, and since $\mathbf{z}$ has an even number of entries, $z_k = -z_{k+1}$. By Lemma 4.4, it is not possible for both z_k and z_{k+1} to be zero, so z_k is equal to some nonzero value c, and this edge contributes $4c^2$ to the value of $\mathbf{z}^T B\mathbf{z}$. On the other hand, there are two edges in G_k that contribute nonzero values and that do not have corresponding edges in G: (v_k, v_{k+1}) and (v_{3k}, v_{3k+1}). Each of these edges contributes c^2 to $\mathbf{x}^T B_k \mathbf{x}$. Thus,

$$\mathbf{x}^T B_k \mathbf{x} = \mathbf{z}^T B\mathbf{z} - 4c^2 + 2c^2 < \mathbf{z}^T B\mathbf{z}.$$

Since $\mathbf{x}^T\mathbf{x} = \mathbf{z}^T\mathbf{z}$,

$$\lambda_2(G_k) \le \frac{\mathbf{x}^T B_k \mathbf{x}}{\mathbf{x}^T\mathbf{x}} < \frac{\mathbf{z}^T B\mathbf{z}}{\mathbf{z}^T\mathbf{z}} = 4\sin^2\left(\frac{\pi}{4k}\right).$$

That is, the second smallest eigenvalue of B_k is less than any nonzero even eigenvalue and is thus odd by Theorem 4.1.

We still need to show that there are not too many zero entries in $\mathbf{u}_2$ (spectral bisection as defined in this paper does not separate vertices with the same value). Since $\mathbf{u}_2$ is an odd vector and since the odd component of G_k is a weighted path graph, Lemmas 4.4 and 5.1 imply that $\mathbf{u}_2$ cannot have consecutive zeros, and the values corresponding to vertices $2k$ and $4k$ are nonzero. Thus the edge separator generated by spectral bisection must cut at least half the edges between the upper and lower paths; since none of these edges share an endpoint, the cover used in generating the vertex separator must include at least this number of vertices. □

Recently Spielman and Teng [27] have presented an algorithm that recursively applies a spectral separator method to give bisections of planar graphs guaranteed to be $O(\sqrt{(n)})$; their technique applied to the roach graph gives a bisection of constant size [27]. See section 8 for details.

6. A bad family of graphs for the "best threshold cut" algorithm. While the roach graph defeats spectral bisection, the second smallest eigenvector can still be used to find a small separator using the "best threshold cut" algorithm. In particular, Theorem 3.1 implies that considering all threshold cuts induced by $\mathbf{u}_2$ produces a constant-size cut: if q_{min} is the minimum cut quotient for these cuts, then

$$q_{min} \le \sqrt{\lambda_2(2\Delta - \lambda_2)} \le \frac{\sqrt{6}\pi}{4k},$$

which implies that q_{min} is $O(\frac{1}{n})$. Since the denominator of q_{min} is less than or equal to $\frac{n}{2}$, the number of edges in this cut must be bounded by a constant.

In this section we show that there is a family of graphs for which the "best threshold cut" algorithm does poorly. The graphs in this family consist of crossproducts of path graphs and double trees. A **double tree** is two complete binary trees of k levels for some $k > 0$ connected by an edge between their respective roots.

The following two bounds are proved in [16].

LEMMA 6.1. *For a complete balanced binary tree on $k \geq 3$ levels and $n = 2^k - 1$ vertices, $\frac{1}{n} < \lambda_2 < \frac{2}{n}$.*

For double trees where each of the component trees has k levels, $n = 2^{k+1} - 2$. The following bound applies.

LEMMA 6.2. *For a double tree on $n \geq 14$ vertices, $\frac{1}{n} < \lambda_2 < \frac{4}{n}$.*

The **tree-cross-path graph** consists of the crossproduct of a double tree on p_1 vertices and a path graph on p_2 vertices. We show that there are tree-cross-path graphs that defeat the "best threshold cut" algorithm.

We formally state the result for this section as follows.

THEOREM 6.3. *There exists a graph G for which the "best threshold cut" algorithm finds a separator S such that the cut quotient for S is bigger than $i(G)$ by a factor as large (to within a constant) as allowed by the bounds from Theorem* 3.1.

Proof. Let G be the tree-cross-path graph that is the crossproduct of a double tree of size p and a path of length $cp^{\frac{1}{2}}$ for some c in the range $3.5 \leq c < 4$. To ensure that the double tree and the path have integer sizes, restrict p to integers of the form $2^k - 2$ for $k \geq 4$. Then choose c in the range specified such that $cp^{\frac{1}{2}}$ is an integer (the choice of p ensures there is an integer in this range).

Recall that the eigenvalues of a graph crossproduct are all pairwise sums of the eigenvalues from the graphs used in the crossproduct operation. Let ν_2 be the second smallest eigenvalue of the double tree on p vertices, and let μ_2 be the second smallest eigenvalue for the path on $cp^{\frac{1}{2}}$ vertices. If $\mu_2 < \nu_2$, then λ_2 for the crossproduct is μ_2 (i.e., μ_2 added to the zero eigenvalue of the double tree). Since $\mu_2 = 4\sin^2(\pi/2cp^{\frac{1}{2}})$ and $\nu_2 \geq \frac{1}{p}$ (by Lemma 6.2 and the choice of p), it is necessary to show that $4\sin^2(\pi/2cp^{\frac{1}{2}}) < \frac{1}{p}$. Reorganizing, simplifying, and noting that $\sin(\theta) < \theta$ for $0 < \theta \leq \frac{\pi}{2}$, it is sufficient to show that $\pi < c$. Clearly by the choice of c, this inequality holds.

Note that the tree-cross-path graph can be thought of as $cp^{\frac{1}{2}}$ copies of the double tree, each corresponding to one vertex of the path graph. Each vertex in the ith copy of the double tree is connected by an edge to the corresponding vertex in copies $i - 1$ and $i + 1$. This description allows one to construct the eigenvector for the second smallest tree-cross-path eigenvalue as follows. Assign each vertex in double tree copy i the value for vertex i in the path graph eigenvector for μ_2. Note that this is the only possible eigenvector, since path graph eigenvalues are simple by Lemma 4.5.

Now consider any copy of the double tree: every vertex in that copy gets the same value in the characteristic valuation. Thus, the cut S made by the "best threshold cut" algorithm must separate at least two copies of the double tree and thus must cut at least p edges. There is a bisection S^* of size $cp^{\frac{1}{2}}$ (cut the edge between the roots in each double tree); because this cut is a bisection, the ratio between the cut quotient q for S and $i(G)$ is at least as large as the ratio between the sizes of these cuts:

$$\frac{q}{i(G)} \geq \frac{|S|}{|S^*|} \geq \frac{p}{cp^{\frac{1}{2}}} = \Omega\left(p^{\frac{1}{2}}\right).$$

From Theorem 3.1,

$$\frac{\lambda_2}{2} \leq i(G) \leq q \leq \sqrt{\lambda_2(2\Delta - \lambda_2)}.$$

This, plus the fact that the tree-cross-path graph has bounded degree ($\Delta = 5$), implies that

$$\frac{q}{i(G)} \leq \frac{2\sqrt{\lambda_2(2\Delta - \lambda_2)}}{\lambda_2} = O\left(\frac{1}{\sqrt{\lambda_2}}\right) = O\left(p^{\frac{1}{2}}\right).$$

These two bounds imply that, to within a constant factor, the ratio is as large as possible, and the theorem holds. □

7. A bad family of graphs for generalized spectral algorithms.

7.1. Purely spectral algorithms. In section 2 we noted that many variations of spectral partitioning have been suggested. In this section we extend the results of the previous section to cover those variations and many other possibilities, including algorithms that use some number k (where k might depend on n) of the eigenvectors corresponding to the k smallest nonzero eigenvalues. In particular, consider algorithms that meet the following restrictions:

- The algorithm computes a value for each vertex using only the eigenvector components for that vertex from k eigenvectors corresponding to the smallest nonzero eigenvalues (for convenience, we refer to these as the k **smallest eigenvectors**). The function computed can be arbitrary as long as its output depends only on these inputs.
- The algorithm partitions the graph by choosing some threshold t, and then putting all vertices with values greater than t on one side of the partition and the rest of the vertices on the other side.
- The algorithm is free to compute the break point t in any way; e.g., checking the cut quotient for all possible breaks and choosing the best one is allowed.

We call such an algorithm **purely spectral**.

7.2. Purely spectral algorithms that use a constant number of eigenvectors. The following theorem gives a bound on how well such algorithms do when the number of eigenvectors used is a constant.

THEOREM 7.1. *Consider the purely spectral algorithms that use the k smallest eigenvectors for k a fixed constant. Then there exists a family of graphs $\mathcal{G}$ such that $G \in \mathcal{G}$ has a bisection S^* with $|S^*| \geq (k^2 n)^{\frac{1}{3}}$ and such that any purely spectral algorithm using the k smallest eigenvectors produces a separator S for G such that $|S| \geq \left(\frac{|S^*|}{\pi k+1}\right)^2$.*

Proof. We show that $\mathcal{G}$ is the set of tree-cross-path graphs that are the crossproducts of double trees of size p (where p is an integer of the form $2^j - 2$ for some $j \geq 4$) and paths of length $cp^{\frac{1}{2}}$, where c is a constant chosen such that $\pi k < c \leq \pi k + 1$ and $cp^{\frac{1}{2}}$ is an integer.

Using slight modifications of arguments from Theorem 6.3, one can show the following. For graphs in $\mathcal{G}$, the k smallest positive path eigenvalues are less than the smallest positive eigenvalue of the double tree. This implies that every vertex in a particular copy of the double tree receives the same set of values from the k eigenvectors. Thus the purely spectral algorithm assigns the same value to each vertex in that copy. This implies that S, the separator produced, must separate at least two copies of the double tree and thus must cut at least p edges.

There is a bisection S^* of size $cp^{\frac{1}{2}}$ (it cuts the edge between the roots in each double tree); because $n = cp^{\frac{3}{2}}$ and $c > k$, it is the case that $|S^*| > k^{\frac{2}{3}} n^{\frac{1}{3}}$. It is obvious that

$$|S| \geq \left(\frac{|S^*|}{c} \right)^2;$$

since $c \leq \pi k + 1$, the theorem holds. □

Note that for the case in which k is constant, the following results apply for the family of graphs described in the preceding theorem:

- The cut quotient q_S is no better than the best cut quotient q_{min} produced by considering all threshold cuts for $\mathbf{u}_2$, and
- the gap between $i(G)$ and q_{min} is as large as possible (within a constant factor) with respect to Theorem 3.1. The bound on $|S^*|$ implies that the spectral separator is bigger by a factor of at least a constant times $n^{\frac{1}{3}}$.

These results can be shown using techniques from the previous section. Thus, for such graphs, using k eigenvectors does not improve the performance of the "best threshold cut" algorithm.

These results also hold for certain variants of the definition of "purely spectral." For example, Chan, Gilbert, and Teng have proposed using the entries of eigenvectors 2 through $d+1$ of the Laplacian as spatial coordinates for the corresponding vertices of a graph [7]. The graph is then partitioned using a geometric separator algorithm [21], [15]. If this technique is applied (using a fixed d) to the counterexample graph used in the proof above, all vertices in a particular copy of the double tree end up with the same coordinates; the geometric algorithm then cuts between copies of the double tree, yielding the same bad cuts as in the proof.

7.3. Purely spectral algorithms that use more than a constant number of eigenvectors. There are still a number of open questions about the performance of purely spectral algorithms that use more than a constant number of eigenvectors (in particular, how well can such algorithms do if they use all the eigenvectors?). However, just using more than a constant number of eigenvectors is not sufficient to guarantee good separators. In particular, the counterexamples and arguments in the previous sections can be extended to prove the following theorem.

THEOREM 7.2. *For sufficiently large n and $0 < \epsilon < \frac{1}{4}$, there exists a bounded-degree graph G on n vertices such that any purely spectral algorithm using the n^ϵ smallest eigenvectors produces a separator S for G with a cut quotient greater than $i(G)$ by at least a factor of $n^{\left(\frac{1}{4}-\epsilon\right)} - 1$.*

Proof. Once again, let G be the tree-cross-path graph. As in the previous two proofs, choose p_1 (the double-tree size) and p_2 (the path size) such that the smallest n^ϵ eigenvalues of the crossproduct are the same as the smallest n^ϵ eigenvalues of the path graph. Once again, a purely spectral algorithm separates two adjacent double trees, while the edges between the roots of the double trees form a better separator. It remains to choose p_1 and p_2 such that the claim about the smallest eigenvalues of the crossproduct holds and to show that the resulting cut is bad.

Set p_1 to some arbitrary positive integer p, subject to the conditions presented below to ensure that p is sufficiently large. Then set $p_2 = \left\lceil p^{\left(\frac{1}{2}+2\epsilon\right)} \right\rceil$. Note that p can be chosen sufficiently large such that

$$p > p^{\left(\frac{1}{2}+2\epsilon\right)} + 1 > p_2.$$

This implies that $p > n^{\frac{1}{2}}$, where $n = p_1 p_2$. Note that this allows one to show easily that $n^\epsilon < p^{2\epsilon} < p_2$ (i.e., since the algorithm uses n^ϵ eigenvectors, this argument requires the path graph to have at least that many eigenvalues and thus be at least that long). Also note that even for fairly small p, $p_2 < 2p^{\left(\frac{1}{2}+2\epsilon\right)}$, which implies that

$$n < 2p^{\left(\frac{3}{2}+2\epsilon\right)}. \tag{7.1}$$

Now consider the ratio of the size of the cut produced by cutting the double-tree edges to the size of the cut produced by a purely spectral method under the assumption that the n^ϵ smallest eigenvalues are the same as for the path graph. As in previous proofs, this ratio is at least as large as the ratio between the number of edges cut. Thus, for sufficiently large p, the ratio is at least

$$\frac{p}{\left\lceil p^{\left(\frac{1}{2}+2\epsilon\right)} \right\rceil} > p^{\left(\frac{1}{2}-2\epsilon\right)} - 1 > n^{\left(\frac{1}{4}-\epsilon\right)} - 1.$$

All that is left to prove is the assumption about the smallest eigenvalues. If $\alpha = \frac{1}{2} - 2\epsilon$, then $\alpha > 0$ and inequality (7.1) above can be written as

$$n < 2p^{(2-\alpha)}. \tag{7.2}$$

Recall that the eigenvalues of a path graph on k vertices are $4\sin^2(\frac{\pi i}{2k})$ for $0 \le i < k$ and that λ_2 for a double tree on $p \ge 14$ vertices is greater than or equal to $\frac{1}{p}$. It remains to show that for p sufficiently large,

$$4\sin^2\left(\frac{\pi n^\epsilon}{2\left\lceil p^{\left(\frac{1}{2}+2\epsilon\right)}\right\rceil}\right) < \frac{1}{p}.$$

Reorganizing, simplifying, noting that $\sin(\theta) < \theta$ for $0 < \theta \le \frac{\pi}{2}$, and applying inequality (7.2) above, it is sufficient to show that there is a sufficiently large p such that

$$\frac{\pi\left(2p^{(2-\alpha)}\right)^\epsilon}{2p^{\left(\frac{1}{2}+2\epsilon\right)}} < \frac{1}{2p^{\frac{1}{2}}}, \quad \text{or} \quad \pi 2^\epsilon < p^{\alpha\epsilon}.$$

Clearly this inequality holds for sufficiently large p. ∎

8. A note on more recent developments. Subsequent to the initial appearance of these results [17], Spielman and Teng published a paper on the performance of spectral partitioning algorithms [28]. Their work has several parts, including

- a proof that for any bounded-degree planar graph, $\lambda_2 = O(n^{-1})$, and that for well-shaped meshes in d dimensions, $\lambda_2 = O(n^{-\frac{2}{d}})$.
- a new proof of a theorem credited to Mihail [20] that extends bounds on quotient cuts to all vectors with small Rayleigh quotients.
- a recursive spectral bisection algorithm. The algorithm produces $O(n^{\frac{1}{2}})$ bisectors for planar graphs and $O(n^{1-\frac{1}{d}})$ bisectors for well-shaped d-dimensional meshes.
- a new bounded-degree planar counterexample graph for which "best threshold cut" gives a poorly balanced separator.

It is interesting to consider how those results relate to the results in this paper.

We have shown that there are bounded-degree planar graphs for which spectral bisection based on $\mathbf{u}_2$ alone gives a cut of size $\Theta(n)$. Spielman and Teng's recursive spectral bisection algorithm, however, produces constant-size bisections for our counterexamples. Thus their algorithm gives a greatly improved, if somewhat more expensive, result. Their bounded-degree counterexample graph is an interesting advance over the roach graph in that it gives a bounded-degree planar graph with both a bad bisection and a poorly balanced "best threshold cut."

As for the tree-cross-path examples, the two papers illustrate the difference between guarantees on the size of a balanced cut versus its optimality. If, on the first cut, the recursive algorithm produces a bisection that is large relative to the best bisection, the recursion will not improve the bisection. (This is the case for the tree-cross-path graph.) Examples exist for well-shaped meshes. The following graph was suggested by John Gilbert. Let a double grid be a pair of $k \times k$ square grids that share a single common corner. As shown in [19], λ_2 of the double grid is $\Theta(\frac{1}{k^2 \log k})$. The double-grid-cross-path graph is a crossproduct between a double grid graph and a path graph. Note that for a suitable constant c, if the path has length $ck\sqrt{\log k}$, the path graph contributes the second smallest eigenvalue of the double grid cross path. Following an analysis similar to that in Theorem 6.3, one can show that the "best threshold cut" for such a double grid cross path is a bisection of size $\Theta(k^2)$ that splits the graph between two copies of the double grid. It is easy to check that the recursive algorithm also returns this cut and thus does not improve the quality of the single spectral cut. However, this example has a bisection of size $\Theta(k\sqrt{\log k})$ (separate the grids at their common points). The larger bisection meets the guarantee for three-dimensional grids (n here is $ck^3\sqrt{\log k}$), but is not optimum.

Acknowledgments. Special thanks are due to Dafna Talmor and Doug Tygar who provided valuable comments on earlier versions of this paper.

REFERENCES

[1] N. Alon, *Eigenvalues and expanders*, Combinatorica, 6 (1986), pp. 83–96.
[2] N. Alon, Z. Galil, and V. D. Milman, *Better expanders and superconcentrators*, J. Algorithms, 8 (1987), pp. 337–347.
[3] N. Alon and V. D. Milman, λ_1*, isoperimetric inequalities for graphs, and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73–88.
[4] B. Aspvall and J. R. Gilbert, *Graph coloring using eigenvalue decomposition*, SIAM J. Algebraic Discrete Meth., 5 (1984), pp. 526–538.
[5] E. R. Barnes, *An algorithm for partitioning the nodes of a graph*, SIAM J. Algebraic Discrete Meth., 3 (1982), pp. 541–550.
[6] R. Boppana, *Eigenvalues and graph bisection: An average-case analysis*, in Proc. 28th Annual IEEE Symposium on Foundations of Computer Science, Los Angeles, October 1987, IEEE Computer Society Press, Los Alamitos, CA, pp. 280–285.
[7] T. F. Chan, J. R. Gilbert, and S. H. Teng, *Geometric spectral partitioning*, Tech. report CSL-94-15, Xerox PARC, Palo Alto, CA, July 1994. Revised January 1995.
[8] D. M. Cvetković, M. Doob, and H. Sachs, *Spectra of Graphs*, Academic Press, New York, 1979.
[9] W. E. Donath and A. J. Hoffman, *Lower bounds for the partitioning of graphs*, IBM J. Res. Develop., 17 (1973), pp. 420–425.
[10] J. Falkner, F. Rendl, and H. Wolkowicz, *A computational study of graph partitioning*, Math. Programming, 66 (1994), pp. 211–240.
[11] M. Fiedler, *Algebraic connectivity of graphs*, Czechoslovak Math. J., 23 (1973), pp. 298–305.
[12] M. Fiedler, *A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory*, Czechoslovak Math. J., 25 (1975), pp. 619–633.

[13] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP–Completeness*, W. H. Freeman, San Francisco, 1979.

[14] M. R. Garey, D. S. Johnson, and L. Stockmeyer, *Some simplified NP-complete graph problems*, Theoret. Comput. Sci., 1 (1976), pp. 237–267.

[15] J. Gilbert, G. Miller, and S. H. Teng, *Geometric mesh partitioning: Implementation and experiments*, in Proc. 9th International IEEE Symposium on Parallel Processing, Santa Barbara, CA, April 1995, IEEE Computer Society Press, Los Alamitos, CA.

[16] S. Guattery and G. Miller, *On the performance of spectral graph partitioning methods*, Tech. report CMU-CS-94-228, Carnegie Mellon University, Pittsburgh, PA, December 1994.

[17] S. Guattery and G. Miller, *On the performance of spectral graph partitioning methods*, in Proc. 6th Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, January 1995, SIAM, Philadelphia, PA, pp. 233–242.

[18] L. Hagen and A. B. Kahng, *New spectral methods for ratio cut partitioning and clustering*, IEEE Trans. Computer-Aided Design, 11 (1992), pp. 1074–1085.

[19] N. Kahale, *A semidefinite bound for mixing rates of Markov chains*, in Proc. 5th Annual Integer Programming and Combinatorial Optimization Conference, Lecture Notes in Comput. Sci. 1084, Springer-Verlag, New York, 1996, pp. 190–203.

[20] M. Mihail, *Conductance and convergence of Markov chains—a combinatorial treatment of expanders,* in Proc. 30th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, 1989, pp. 526–531.

[21] G. L. Miller, S. H. Teng, and S. A. Vavasis, *A unified geometric approach to graph separators*, in Proc. 32nd Annual IEEE Symposium on Foundations of Computer Science, Puerto Rico, October 1991, IEEE Computer Society Press, Los Alamitos, CA, pp. 538–547.

[22] B. Mohar, *The Laplacian spectrum of graphs*, in Graph Theory, Combinatorics, and Applications, Vol. 2, Wiley Interscience, New York, pp. 871–898.

[23] B. Mohar, *Isoperimetric numbers of graphs*, J. Combin. Theory Ser. B, 47 (1989), pp. 274–291.

[24] A. Pothen, H. D. Simon, and K. P. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.

[25] F. Rendl and H. Wolkowicz, *A projection technique for partitioning the nodes of a graph*, Ann. Oper. Res., 58 (1995), pp. 155–180.

[26] H. D. Simon, *Partitioning of unstructured problems for parallel processing*, Comput. Systems Engrg., 2 (1991), pp. 135–148.

[27] A. Sinclair and M. Jerrum, *Approximate counting, uniform generation and rapidly mixing Markov chains*, Inform. Comput., 82 (1989), pp. 93–133.

[28] D. A. Spielman and S. H. Teng, *Spectral partitioning works: Planar graphs and finite element meshes*, Tech. report UCB CSD-96-898, University of California, Berkeley, CA, 1996, an extended abstract also appears in Proc. 37th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computer Society Press, Los Alamitos, CA, 1996, pp. 96–105.

[29] D. M. Young and R. T. Gregory, *A Survey of Numerical Mathematics*, Addison–Wesley Series in Mathematics, Vol. 2, Addison–Wesley, Reading, MA, 1973.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 720–736, July 1998

008

STABILITY OF CONJUGATE GRADIENT AND LANCZOS METHODS FOR LINEAR LEAST SQUARES PROBLEMS*

ÅKE BJÖRCK†, TOMMY ELFVING†, AND ZDENĚK STRAKOŠ‡

Abstract. The conjugate gradient method applied to the normal equations $A^TAx = A^Tb$ (CGLS) is often used for solving large sparse linear least squares problems. The mathematically equivalent algorithm LSQR based on the Lanczos bidiagonalization process is an often recommended alternative. In this paper, the achievable accuracy of different conjugate gradient and Lanczos methods in finite precision is studied. It is shown that an implementation of algorithm CGLS in which the residual $s_k = A^T(b - Ax_k)$ of the normal equations is recurred will not in general achieve accurate solutions. The same conclusion holds for the method based on Lanczos bidiagonalization with starting vector A^Tb. For the preferred implementation of CGLS we bound the error $\|r - r_k\|$ of the computed residual r_k. Numerical tests are given that confirm a conjecture of backward stability. The achievable accuracy of LSQR is shown to be similar. The analysis essentially also covers the preconditioned case.

Key words. conjugate gradient method, least squares, numerical stability

AMS subject classification. 65F20

PII. S089547989631202X

1. Introduction.

Iterative methods are useful alternatives to direct methods for several classes of large sparse least squares problems, see [13, 3]. In this paper we compare different implementations of Krylov subspace methods for solving the linear least squares problem

$$\min_x \|b - Ax\|_2, \tag{1.1}$$

where $A \in \mathbf{R}^{m\times n}$, $m \geq n$, is a given matrix. We will assume that $\operatorname{rank}(A) = n$, although some of the conclusions also hold for $\operatorname{rank}(A) < n$.

It is well known that x is a least squares solution if and only if the residual vector $r = b - Ax \perp \mathcal{R}(A)$, or equivalently when $A^T(b - Ax) = 0$. The resulting system of normal equations

$$A^TAx = A^Tb \tag{1.2}$$

is always consistent. The implementations include conjugate gradient methods and methods based on two different versions of Lanczos bidiagonalization.

The conjugate gradient method (CG), developed in the early 1950s, has become a basic tool for solving large sparse linear systems and linear least squares problems. In the original paper by Hestenes and Stiefel [14], and in the subsequent paper [23], a version of the conjugate gradient method for the solution of the normal equations (1.2) was given. Läuchli [17] discussed a preconditioned conjugate gradient method

*Received by the editors November 13, 1996; accepted for publication (in revised form) by A. Greenbaum May 4, 1997; published electronically March 18, 1998.
http://www.siam.org/journals/simax/19-3/31202.html

†Department of Mathematics, Linköping University, S-581 83, Linköping, Sweden (akbjo@math.liu.se, toelf@math.liu.se). The work of T. Elfving was supported by the Swedish Research Council for Engineering Sciences (TFR) under contract 222-95-401.

‡Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 182 07 Praha 8, Czech Republic (strakos@uivt.cas.cz). The work of this author was supported by AS CR grant A2030706.

for solving least squares geodetic network problems. The application of CG methods to linear least squares problems has also been discussed by Lawson [18] and Chen [5]. Paige [19] derived a method LSCG based on the Lanczos bidiagonalization process of Golub and Kahan in [9]. This method was later shown to be numerically unstable, and a stable version called LSQR was given by Paige and Saunders in [20].

Reid [21] gave an excellent discussion of different computational variants of the conjugate gradient method for solving symmetric positive definite systems. In theory, applying the conjugate gradient method to the normal equations is a straightforward extension of the standard conjugate gradient method. However, the actual implementation is critical and numerically unstable variants are still seen in the literature! Although no comprehensive comparison of different implementations has been published, several conclusions can be found in Elfving [7] and in Paige and Saunders [20]. The aim of this paper is to compare the achievable accuracy in finite precision of different implementations of Krylov subspace methods for solving (1.2). We will make heavy use of recent important results on the behavior of Lanczos and conjugate gradient methods in finite precision, see Greenbaum [10, 11], Greenbaum and Strakoš [12], and Strakoš [24]. The finite precision behavior of stationary iterative methods is studied in [16].

In section 2 we give a survey of Krylov subspace methods for solving the normal equations and their properties in exact arithmetic. In section 3 different implementations of conjugate gradient or Lanczos type methods are considered and their computational complexity compared. The performance of the algorithms in finite precision is discussed in sections 4 and 5. In section 4 it is shown that some implementations of conjugate gradient and Lanczos type methods fail to give accurate solutions. The recommended implementations of conjugate gradient and LSQR are analyzed in section 5. An upper bound for the residual error is derived, which shows these implementations to be backward stable. This extends the analysis of Greenbaum [11] to inconsistent least squares problems. Numerical test results are given in section 6 that confirm the theoretical analysis.

2. Krylov space methods for least squares. When A has full rank the system of normal equations (1.2) has a unique solution

$$x = A^{\dagger}b, \qquad A^{\dagger} = (A^TA)^{-1}A^T.$$

We denote by $r = b - Ax$ the corresponding residual. For a given starting vector x_0 the conjugate gradient algorithm generates approximations x_k in the affine subspace

$$x_k \in x_0 + \mathcal{K}_k(A^TA, s_0), \tag{2.1}$$

$s_0 = A^T(b - Ax_0)$, where

$$\mathcal{K}_k(A^TA, s_0) = \operatorname{span}\left\{s_0, (A^TA)s_0, \ldots, (A^TA)^{k-1}s_0\right\} \tag{2.2}$$

is a Krylov subspace. The iterates generated are optimal in the sense that for each k, x_k minimizes the error functional

$$E_\mu(y) = (x-y)^T(A^TA)^\mu(x-y), \qquad y \in x_0 + \mathcal{K}_k(A^TA, s_0). \tag{2.3}$$

Only the values $\mu = 0, 1, 2$ are of practical interest. By (2.3) and using

$$A(x - x_k) = b - r - Ax_k = r_k - r,$$

where $r_k = b - Ax_k$, we obtain

$$E_\mu(x_k) = \begin{cases} \|x - x_k\|^2, & \mu = 0; \\ \|r - r_k\|^2 = \|r_k\|^2 - \|r\|^2, & \mu = 1; \\ \|A^T(r - r_k)\|^2 = \|A^T r_k\|^2, & \mu = 2. \end{cases} \tag{2.4}$$

Here and in the following, $\|\cdot\|$ denotes the l_2-norm. The second expression for $E_1(x_k)$ in (2.4) follows from the fact that $r \perp r - r_k$.

Using the inner product and norm

$$(x, y)_\mu = x^T (A^T A)^\mu y, \qquad \|x\|_\mu^2 = (x, x)_\mu$$

the conjugate gradient method can be formulated as follows.

ALGORITHM 2.1 (CGLS (μ)).

Let x_0 be an intial approximation, put

$$r_0 = b - Ax_0, \quad s_0 = p_1 = A^T r_0, \quad \gamma_0 = \|s_0\|_{\mu-1}^2, \tag{2.5}$$

and for $k = 1, 2, \ldots$ compute

$$\begin{aligned} q_k &= Ap_k, \\ \alpha_k &= \gamma_{k-1}/\|p_k\|_\mu^2, \\ x_k &= x_{k-1} + \alpha_k p_k, \\ r_k &= r_{k-1} - \alpha_k q_k, \\ s_k &= A^T r_k, \\ \gamma_k &= \|s_k\|_{\mu-1}^2, \\ \beta_k &= \gamma_k/\gamma_{k-1}, \\ p_{k+1} &= s_k + \beta_k p_k. \end{aligned} \tag{2.6}$$

For $\mu = 0$ the method is equivalent to Craig's method [6], which is called CGNE in [8]. Note that $\mu = 0$ is feasible only when $b \in \mathcal{R}(A)$, since we need to be able to evaluate $s_k^T (A^T A)^{-1} s_k$, where $s_k = A^T r_k$ is the current residual of the normal equations. If $b \in \mathcal{R}(A)$, then we have

$$s_k^T (A^T A)^{-1} s_k = (b - Ax_k)^T A (A^T A)^{-1} A^T (b - Ax_k) = r_k^T r_k,$$

where $A(A^T A)^{-1} A^T$ is the orthogonal projection onto $\mathcal{R}(A)$, see also [19].

For $\mu > 0$ we use $\|p_k\|_\mu^2 = \|q_k\|_{\mu-1}^2$. For $\mu = 1$ the method is called CGLS in [20] and CGNR in [8].

The variational property of the conjugate gradient method implies that in exact arithmetic the error functional $E_\mu(x_k)$ decreases monotonically as a function of k. By (2.4), $\|r_k\|$ will also decrease monotonically for $\mu = 1$. Further, we have the following result.

LEMMA 2.1. *Let $\{x_k\}$ be the sequence of conjugate gradient approximations, which minimize $E_\mu(y)$ subject to $y \in x_0 + \mathcal{K}_k(A^T A, s_0)$. Then for $\mu = 1, 2$ the sequences $E_\nu(x_k)$, $0 \le \nu \le \mu$ all decrease monotonically as functions of k.*

Proof. For $\mu = 1$ see [14, p. 416]. For $\mu = 2$ see [14, Section 7]. ∎

It follows that for $\mu = 1$ both $\|r - r_k\|$ and $\|x - x_k\|$ decrease monotonically. However, $\|A^T r_k\|$ will often exhibit large oscillations when $\kappa(A)$ is large. We stress that this behavior is *not* a result of rounding errors. For $\mu = 2$ $\|A^T r_k\|$ will also

decrease monotonically, but this choice gives slower convergence in $\|r - r_k\|$ and $\|x - x_k\|$. It also gives lower final accuracy in these quantities, see [1], and requires more operations or storage. Therefore we consider here only the case $\mu = 1$, which is of most practical interest.

From the optimality property follows the upper bound on the rate of convergence

$$E_\mu(x_k) \le 2\left(\frac{\kappa - 1}{\kappa + 1}\right)^k E_\mu(x_0), \tag{2.7}$$

where $\kappa = \kappa(A) = \sqrt{\kappa(A^T A)}$, see [3, Chapter 7.4]. However, the convergence of the conjugate gradient method also depends on the distribution of the singular values of A. In particular, if A has only $t \le n$ distinct singular values, then in exact arithmetic the solution is obtained in at most t steps.

Often a right preconditioner S is used, which corresponds to the transformation

$$\min_y \|(AS^{-1})y - b\|_2, \qquad Sx = y. \tag{2.8}$$

Typically S is chosen as an approximation of the Cholesky factor R of $A^T A$. If $S = R$, the matrix AS^{-1} is orthogonal and the conjugate gradient method converges in one iteration.

Although (2.7) essentially continues to also hold in finite precision, this is not true of the finite termination property, and conjugate gradient methods should be regarded as iterative methods. In many practical applications, however, one is satisfied with approximations that are obtained in far less than n iterations.

3. Implementation. There are many ways, all mathematically equivalent, in which to implement the conjugate gradient method. In *exact* arithmetic they will all generate the same sequence of approximations, but in finite precision the achieved accuracy may differ substantially. It is important to notice that an implementation of the conjugate gradient method for symmetric positive definite systems should not be applied directly to the normal equations. In particular the explicit formation of the matrix $A^T A$ should be avoided. All the algorithms below require two matrix vector multiplications of the form Ap and $A^T q$ per iteration step.

3.1. The conjugate gradient method CGLS. The algorithm originally given by Hestenes and Stiefel [14, p. 424] and Stiefel [23] is as follows.

ALGORITHM 3.1 (CGLS1).

Set

$$r_0 = b - Ax_0, \quad s_0 = p_1 = A^T r_0, \quad \gamma_0 = \|s_0\|^2, \tag{3.1}$$

and for $k = 1, 2, \ldots$ *compute*

$$\begin{aligned} q_k &= Ap_k, \\ \alpha_k &= \gamma_{k-1}/\|q_k\|^2, \\ x_k &= x_{k-1} + \alpha_k p_k, \\ r_k &= r_{k-1} - \alpha_k q_k, \\ s_k &= A^T r_k, \\ \gamma_k &= \|s_k\|^2, \\ \beta_k &= \gamma_k/\gamma_{k-1}, \\ p_{k+1} &= s_k + \beta_k p_k. \end{aligned} \tag{3.2}$$

Elfving [7] compared CGLS1 with several other implementations of the conjugate gradient method and found this to be the most accurate. CGLS1 requires storage of two n-vectors x, p and two m vectors r, q. (Note that s can share storage with q.) Each iteration requires about $2\mathrm{nz}(A) + 2m + 3n$ multiplications, where $\mathrm{nz}(A)$ is the number of nonzero elements in A.

A small variation of Algorithm CGLS1 is obtained if instead of r_k the residual of the normal equations $s = A^T(b - Ax)$ is recurred.

ALGORITHM 3.2 (CGLS2).

Let x_0 be an initial approximation, set

$$s_0 = p_1 = A^T(b - Ax_0), \quad \gamma_0 = \|s_0\|^2, \tag{3.3}$$

and for $k = 1, 2, \ldots$ compute

$$\begin{aligned} q_k &= Ap_k, \\ \alpha_k &= \gamma_{k-1}/\|q_k\|^2, \\ x_k &= x_{k-1} + \alpha_k p_k, \\ s_k &= s_{k-1} - \alpha_k(A^T q_k), \\ \gamma_k &= \|s_k\|^2, \\ \beta_k &= \gamma_k/\gamma_{k-1}, \\ p_{k+1} &= s_k + \beta_k p_k. \end{aligned} \tag{3.4}$$

CGLS2 requires the storage of three n-vectors x, p, s and one m vector q and $2\mathrm{nz}(A) + 4n + m$ multiplications.

3.2. Methods based on Lanczos bidiagonalization. Paige and Saunders [20] developed algorithms based on the Lanczos bidiagonalization process of Golub and Kahan [9]. There are two forms of this bidiagonalization procedure, Bidiag1 and Bidiag2, that produce two algorithms that differ in their numerical properties.

In Bidiag1 we start the recursion with $\beta_1 u_1 = b - Ax_0$, $\beta_1 = \|b - Ax_0\|$. After k steps we have computed

$$V_k = (v_1, \ldots, v_k), \qquad U_{k+1} = (u_1, \ldots, u_{k+1}),$$

$$B_k = \begin{pmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \beta_3 & \ddots & \\ & & \ddots & \alpha_k \\ & & & \beta_{k+1} \end{pmatrix} \in \mathbf{R}^{(k+1)\times k}. \tag{3.5}$$

In exact arithmetic it holds that $AV_k = U_{k+1}B_k$, $V_k^T V_k = I_k$, $U_{k+1}^T U_{k+1} = I_{k+1}$. The columns of V_k form an orthonormal basis for the Krylov subspace $\mathcal{K}_k(A^T A, s_0)$. If we set $x_k - x_0 = V_k y_k$, then $\|b - Ax_k\|$ is minimized when y_k solves the linear least squares problem

$$\min_{y_k} \|B_k y_k - \beta_1 e_1\|.$$

This problem can easily be solved by reducing B_k to upper bidiagonal form by a sequence of Givens rotations. This leads to the LSQR algorithm in [20].

ALGORITHM 3.3 (LSQR (Bidiag1)).
Initialize

$$\beta_1 u_1 = b - Ax_0, \quad \alpha_1 v_1 = A^T u_1, \quad w_1 = v_1,$$
$$\bar{\phi}_1 = \beta_1, \quad \bar{\rho}_1 = \alpha_1,$$

and for $k = 1, 2, \ldots$ *repeat*

$$\begin{aligned}
&\beta_{k+1} u_{k+1} = Av_k - \alpha_k u_k,\\
&\alpha_{k+1} v_{k+1} = A^T u_{k+1} - \beta_{k+1} v_k,\\
&[c_k, s_k, \rho_k] = \text{givrot}(\bar{\rho}_k, \beta_{k+1}),\\
&\theta_{k+1} = s_k \alpha_{k+1}, \quad \bar{\rho}_{k+1} = c_k \alpha_{k+1},\\
&\phi_k = c_k \bar{\phi}_k, \quad \bar{\phi}_{k+1} = -s_k \bar{\phi}_k,\\
&x_k = x_{k-1} + (\phi_k/\rho_k) w_k,\\
&w_{k+1} = v_{k+1} - (\theta_{k+1}/\rho_k) w_k.
\end{aligned}$$

Here the scalars $\alpha_i \geq 0$ and $\beta_i \geq 0$ are chosen to normalize the vectors v_i and u_i, respectively, and $[c, s, \sigma] = \text{givrot}(\alpha, \beta)$ is a subroutine that computes c, s, σ in a Givens rotations such that $-s\alpha + c\beta = 0$ and $\sigma = (\alpha^2 + \beta^2)^{1/2}$.

In addition to the $2\text{nz}(A)$ multiplications required by all versions, LSQR requires $3m+5n$ multiplications and storage of two m-vectors u, Av and three n-vectors x, v, w. A basic relation between LSQR and CGLS1 is that v_i and w_i are proportional to s_{i-1} and p_i. Paige and Saunders showed in [20] by numerical examples that LSQR tends to converge slightly faster than CGLS1 when A is ill-conditioned.

Paige derived in [19] an analytically equivalent algorithm LSCG using a variant of the Lanczos bidiagonalization called Bidiag2. In Bidiag2 the Lanczos recursion is started with $\theta_1 v_1 = A^T(b - Ax_0)$, $\theta_1 = \|A^T(b - Ax_0)\|$. After k steps we have computed

$$V_k = (v_1, \ldots, v_k), \qquad P_k = (p_1, \ldots, p_k),$$

$$R_k = \begin{pmatrix} \rho_1 & \theta_2 & & & \\ & \rho_2 & \theta_3 & & \\ & & \ddots & \ddots & \\ & & & \rho_{k-1} & \theta_k \\ & & & & \rho_k \end{pmatrix} \in \mathbf{R}^{k \times k}, \tag{3.6}$$

and in exact arithmetic it holds that $AV_k = P_k R_k$, $V_k^T V_k = P_k^T P_k = I_k$. Here R_k is already in upper triangular form. Hence $x_k - x_0 = V_k y_k$, where y_k is obtained from $R_k^T R_k y_k = \theta_1 e_1$, and the method is implemented in the form

$$R_k^T f_k = \theta_1 e_1, \qquad x_k = (V_k R_k^{-1}) f_k.$$

ALGORITHM 3.4 (LSCG (Bidiag2)).

$$\theta_1 v_1 = A^T(b - Ax_0), \quad \rho_1 p_1 = Av_1, \quad w_1 = v_1/\rho_1, \quad \zeta_0 = -1,$$

TABLE 3.1
Comparison of storage and operations.

	μ	Storage	mults/step
CGLS1:	1	$2n+2m$	$3n+2m$
CGLS2:	1	$3n+m$	$4n+m$
LSQR:	1	$3n+2m$	$5n+3m$
LSCG:	1	$3n+m$	$5n+3m$

and for $k = 1, 2 \ldots$ *compute*

$$\begin{aligned}
&\zeta_k = -(\theta_k/\rho_k)\zeta_{k-1},\\
&x_k = x_{k-1} + \zeta_k w_k,\\
&\theta_{k+1} v_{k+1} = A^T p_k - \rho_k v_k,\\
&\rho_{k+1} p_{k+1} = A v_{k+1} - \theta_{k+1} p_k,\\
&w_{k+1} = (v_{k+1} - \theta_{k+1} w_k)/\rho_{k+1}.
\end{aligned}$$

Here the matrix $V_k = (v_1, \ldots, v_k)$ is the same as in LSQR. LSCG requires $3m + 5n$ multiplications and storage of one m-vector p and three n-vectors x, v, w.

3.3. Storage and operations. All the Krylov methods described above require two matrix vector multiplications costing $2\mathrm{nz}(A)$ multiplications at each iteration step. In the preconditioned case two linear systems of the form $St = p$ and $S^T s = r$ must also be solved at each step. Storage may be needed for A and S. These costs often dominate the total storage and work.

A comparison of additional storage and operations needed for the methods considered is given in Table 3.1. Here CGLS1 shows an advantage over LSQR. Note however that this may be partly offset by the fact that viable rules for stopping the iterations are more costly for CGLS1 than for LSQR.

For LSQR Paige and Saunders [20] consider several stopping rules which require (estimates of) $\|r_k\|, \|x_k\|, \|s_k\|, \|A\|$, and $\|A^\dagger\|$. They show that all these quantities can be obtained at minimal cost in LSQR. For CGLS1 $\|s_k\|$ is available but $\|r_k\|$ has to be separately computed, if needed, at an extra cost of m multiplications.

4. CGLS2 and LSCG in finite precision.

4.1. Floating point arithmetic. In the following analysis we assume that the standard model for floating point computation holds; i.e., if x and y are floating point numbers, then

$$fl(x \; op \; y) = (x \; op \; y)(1+\delta), \qquad |\delta| \le u,$$

where u is the unit roundoff. This leads to a bound for the roundoff error in the product of two matrices $A \in \mathbf{R}^{m\times n}$ and $B \in \mathbf{R}^{n\times p}$,

$$|fl(AB) - AB| \le \gamma_n |A|\,|B|, \qquad \gamma_n = nu/(1-nu), \tag{4.1}$$

where the inequalities are to be interpreted componentwise, see Higham [15]. It is always assumed when this result is used that $nu \ll 1$.

It is well known, see [3, Section, 1.4.3], that an approximate solution $\tilde{x}$ to (1.2) computed by a normwise backward stable method for solving (1.1) will satisfy an

error bound of the form ($x \neq 0$)

$$\|x - \tilde{x}\| \leq \gamma_m \kappa_{LS}(A,b)\|x\|, \tag{4.2}$$

$$\kappa_{LS}(A,b) = \kappa(A)\left(1 + \frac{\|A^\dagger\|\,\|r\|}{\|x\|}\right). \tag{4.3}$$

Here $\kappa(A) = \sigma_1(A)/\sigma_n(A)$ is the spectral condition number of A, and κ_{LS} is the condition number for computing x. Note that if $\|r\| < \|x\|/\|A^\dagger\|$, then $\kappa_{LS} < 2\kappa(A)$, although in general a term proportional to $\kappa^2(A)$ is present.

A first-order perturbation bound corresponding to componentwise relative errors satisfying $|\delta A| \leq \omega|A|$, $|\delta b| \leq \omega|b|$ is given by (see [2])

$$|x - \tilde{x}| \leq \omega|A^\dagger|(|A|\,|x| + |b|) + \omega|(A^TA)^{-1}|\,|A^T|\,|r| + O(\omega^2). \tag{4.4}$$

4.2. Failure of CGLS2 and LSCG. It was pointed out by Paige and Saunders in [20, Section, 7.1] that the explicit use of vectors of the form $A^T(Ap)$ as in CGLS2 can lead to poor performance on ill-conditioned systems. Below we give a new simple explanation for the failure of both CGLS2 and LSCG to obtain good accuracy.

The two algorithms CGLS2 and LSCG share the following feature. Assuming that $x_0 = 0$, the only information about the right-hand side b is in the initialization of $p_0 = s_0 = A^Tb$ and $\theta_1 v_1 = A^Tb$, respectively. Note that *no reference to b is made in the iterative phase.* It follows that roundoff errors that occur in computing the vector A^Tb cannot be compensated for later. By (4.1) we have

$$|fl(A^Tb) - A^Tb| \leq \gamma_m|A^T|\,|b|, \qquad \gamma_m = mu/(1-mu), \tag{4.5}$$

and this is almost sharp. The perturbed solution $x+\delta x$ corresponding to $c = fl(A^Tb)$ satisfies

$$A^TA(x+\delta x) = A^Tb + \delta c, \qquad |\delta c| \leq \gamma_m|A^T|\,|b|.$$

Subtracting $A^TAx = A^Tb$ and solving for δx we obtain $\delta x = (A^TA)^{-1}\delta c$, and from this we get the componentwise bound

$$|\delta x| \leq \gamma_m|(A^TA)^{-1}|\,|A^T|\,|b|. \tag{4.6}$$

Hence if $|b| \gg |r|$, which is the case for a nearly consistent system, this error bound is much larger than the second term in the perturbation bound (4.4).

Using norms we obtain

$$\|\delta x\| \leq \gamma_m\|(A^TA)^{-1}\|\,\|A^T\|\,\|b\| = \gamma_m\kappa(A)\|A^\dagger\|\,\|b\|. \tag{4.7}$$

Since b is not used by CGLS2/LSCG it follows that *this initial error cannot be canceled*, and the best error bound we can hope for will include the term given above.

Comparing with the error bound (4.2)–(4.3) we conclude that if

$$\max\left\{\|x\|/\|A^\dagger\|, \|r\|\right\} \ll \|b\|,$$

CGLS2/LSCG can be expected to produce much less than optimal accuracy. Note that since $x = A^\dagger b$ we could theoretically have $\|x\| = \|A^\dagger\|\,\|b\|$. However, this situation very seldomly reflects the properties of least squares solutions. In particular, ill-conditioned problems usually have solutions with much smaller norms: $\|x\| \ll \|A^\dagger\|\,\|b\|$.

Similarly the residual of the perturbed solution may differ from the true residual by as much as

$$\|b - A(x + \delta x) - (b - Ax)\| = \|A\delta x\| \leq \gamma_m \kappa(A)\|b\|.$$

Again, this cannot be canceled in the iterations, and the iterated residual will (in the best case) converge to the perturbed residual. The numerical experiments in section 6 demonstrate this nicely.

The loss of accuracy will occur even if a preconditioner is used. In CGLS2 we then initialize $s_0 = p_1 = S^{-T}(A^T(b - Ax_0))$, where $S^{-T}A^T$ is never explicitly formed. Therefore the same roundoff errors will occur in computing $A^T(b - Ax_0)$, and by the same reasoning as above these error will never be canceled.

We remark that avoiding the loss of information from explicitly multiplying b by A^T is also important when solving discrete ill-posed linear systems. Using a Krylov subspace method, regularization can be achieved by working with the Krylov subspaces span $\{(AA^T)^p b, (AA^T)^{p+1}b, \ldots, (AA^T)^{p+k}b\}$, where p is some positive integer. In [4] an implicit shift restarted Lanczos method (see [22]) is used to implement such a method in a numerically stable way.

5. Error analysis of CGLS1 and LSQR. Greenbaum [11] studies the finite precision implementation of the class of iterative methods in which each step updates the approximate solution x_k and residual r_k using the formulas

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k p_k, \\ r_{k+1} &= r_k - \alpha_k A p_k. \end{aligned} \tag{5.1}$$

Though in LSQR the residual is not recursively updated, we can formally add the update

$$r_k = r_{k-1} - (\phi_k/\rho_k)Aw_k \tag{5.2}$$

to Algorithm 3.3 without making any other changes in it. It follows then that the analysis based on (5.1) can be applied also to the LSQR method. In our experiments we have used LSQR with this additional recursion.

In [11], the matrix A is assumed square nonsingular. It is easy to see that the analysis will also apply, with a simple modification, to the least squares problem with A rectangular and $r = b - Ax$ different from zero. Let x_k, r_k, α_k, and p_k denote from now on the computed values. Then, following [11], we may easily derive the following result (here and in the rest of the paper we assume for simplicity that $x_0 = 0$).

THEOREM 5.1. *The difference between the true residual $b - Ax_k$ and the recursively computed vector r_k satisfies*

$$\begin{aligned} \frac{\|b - Ax_k - r_k\|}{\|A\|\|x\|} \leq u\,[k + 1 + (1 + c + k(10 + 2c))\Theta_k] + u(k + 1)\frac{\|r\|}{\|A\|\|x\|} \\ + O(u^2), \end{aligned} \tag{5.3}$$

where

$$\Theta_k = \max_{j \leq k} \|x_j\|/\|x\|,$$

u is the machine precision, and c depends on the accuracy of the matrix vector multiply (4.5). *If the matrix-vector product is computed in the standard way, then $c = \gamma_m$.*

By using the relation

$$b - Ax_k - r_k = A(x - x_k) + (r - r_k)$$

we obtain the following slight modification of (5.3):

$$\frac{\|A(x-x_k)\|}{\|A\|\|x\|} \le u\,[k+1+(1+c+k(10+2c))\Theta_k] + u(k+1)\frac{\|r\|}{\|A\|\|x\|} + \\ + \frac{\|r-r_k\|}{\|A\|\|x\|} + O(u^2). \tag{5.4}$$

If it can be shown that there is a constant c_1 such that the computed recursive residual r_k satisfies

$$\frac{\|r-r_k\|}{\|A\|\,\|x\|} \le c_1 u + O(u^2), \quad k \ge S, \tag{5.5}$$

then (5.4) gives an upper bound for the accuracy of the ultimately attainable true residual of the computed approximation. Here S denotes the number of iterations needed to reach a steady-state.

Though we offer no formal proof of (5.5) (to our knowledge, no rigorous formal proofs exist even for the practically used variants of conjugate gradient-type methods for solving linear systems), there is overwhelming experimental evidence that justifies our further considerations. In the following we *assume* that (5.5) holds for both CGLS1 and LSQR. From (5.4) and (5.5) we then get

$$\frac{\|A(x-x_S)\|}{\|A\|\|x\|} \le u\,[S+1+c_1+(1+c+S(10+2c))\Theta] + u(S+1)\frac{\|r\|}{\|A\|\|x\|} \\ + O(u^2), \tag{5.6}$$

where $\Theta = \Theta_S$.

We wish to bound the value Θ. If $\|A^\dagger(r_{k-1} - r_k)\|$ reaches the level $O(u)\|x_{k-1}\|$, then by (5.1) the approximate solution x_k remains essentially unchanged (we assume this situation will take place on or before the step S). In exact arithmetic, both CGLS1 and LSQR are equivalent to the conjugate gradient method applied to the normal equations. By Lemma 2.1, the l_2-norm of the error will therefore decrease monotonically in both cases. Using the analogy developed in [10, 12], and arguing similarly as in [11, Section 3.3], one may therefore expect that the relation $\|x - x_k\| \le \|x - x_0\|$, and thus

$$\|x_k\| \le 2\|x\| + \|x_0\|$$

will hold to a close approximation for the computed quantities. Consequently, we obtain the estimate $\Theta \approx 2 + \Theta_0$. Hence, considering $x_0 = 0$,

$$\frac{\|A(x-x_S)\|}{\|A\|\|x\|} \le u\,(3+c_1+21S+2c+4Sc) + u(S+1)\frac{\|r\|}{\|A\|\|x\|} + O(u^2). \tag{5.7}$$

It follows that the ultimately attainable error of the computed approximation x_S for both CGLS1 and LSQR is bounded by

$$\|x - x_S\| \le u\kappa(A)\left[3+c_1+21S+2c+4Sc+(S+1)\frac{\|r\|}{\|A\|\|x\|}\right]\|x\| + O(u^2). \tag{5.8}$$

In the consistent case, (5.7) shows that the residuals corresponding to the computed approximate solution is of the order of the unit roundoff u, and in this sense CGLS1 and LSQR are normwise backward stable. In the inconsistent case, (5.8) shows that CGLS1 and LSQR may compute *more accurate* solutions than a backward stable method. Note, however, that the bound S for the number of iterations needed depends on $\kappa(A)$.

We note that essentially the same formulas hold for the preconditioned case. Then a vector t_k is computed by solving $St_k = p_k$, but this does not alter the form of the recursion formulas (5.1), which become

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k t_k, \\ r_{k+1} &= r_k - \alpha_k A t_k. \end{aligned} \tag{5.9}$$

In general, the preconditioner will accelerate convergence, and hence the steady-state will be reached for a smaller number of steps S.

It should be mentioned that though both CGLS1 and LSQR were shown here to behave in essentially the same way, some differences may occur for very ill-conditioned problems. In fact in our analysis, we use some implicit restrictions on $\kappa(A)$ allowing us to apply the results [10, 11, 12]. Therefore our results do not apply to highly ill-conditioned problems.

6. Numerical results. Numerical tests were performed in MATLAB on a SUN SPARCstation 10 using double precision with unit roundoff $u = 2.2 \cdot 10^{-16}$. The tests are similar to tests run on an IBM 370 using double precision by the first author during a visit to Stanford University in 1979 [1].

The test problems denoted $P(m, n, d, p)$ were taken from Paige and Saunders [20]. The matrix A was constructed by

$$A = Y \begin{pmatrix} D \\ 0 \end{pmatrix} Z^T \in \mathbf{R}^{m \times n}, \quad Y = I - 2yy^T, \quad Z = I - 2zz^T.$$

Here y and z are Householder vectors of appropriate dimension, with elements

$$\begin{aligned} y_i &= \sin(4\pi i/m), \quad i = 1, \ldots, m, \\ z_i &= \cos(4\pi i/n), \quad i = 1, \ldots, n, \end{aligned}$$

followed by normalization so that $\|y\| = \|z\| = 1$. For $n = q$ $(d = 1)$ the singular values are chosen to be

$$D = q^{-p} \mathrm{diag}\,(q^p, \ldots, 3^p, 2^p, 1);$$

i.e., p is a power factor. Taking $n = qd$ leads to d copies of each singular value. The solution is taken to be $x = (n-1, \ldots, 2, 1, 0)^T$, and the right-hand side is constructed as

$$\begin{aligned} b &= Ax + r, \quad r = \rho Y \begin{pmatrix} 0 \\ c \end{pmatrix}, \\ c &= m^{-1}(1, -2, 3, \ldots, \pm(m-n))^T. \end{aligned}$$

Thus if $m > n$ and $\rho > 0$ the system is incompatible and $\|r\| \approx \rho$. (In [20] $\rho = 1$ was used throughout.)

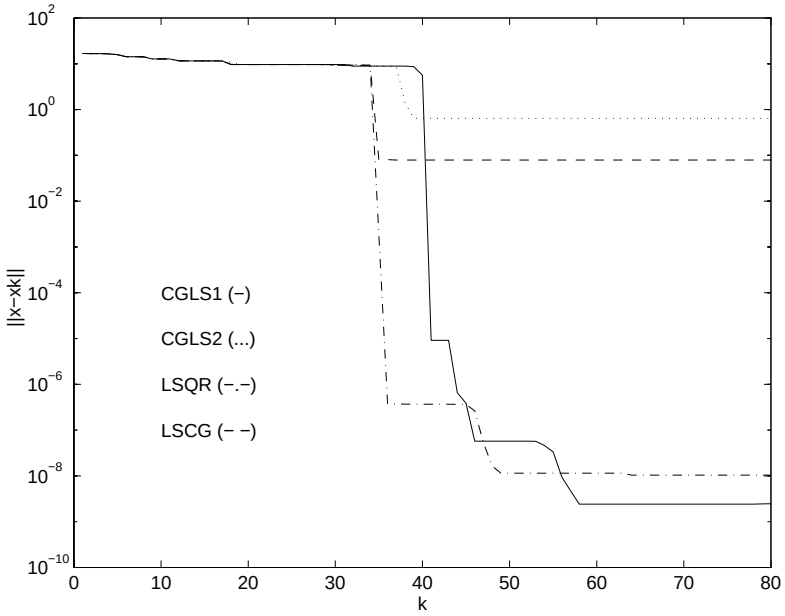


FIG. 6.1. $\|x - x_k\|$ *for problem* $PS(10, 10, 1, 8)$, $\kappa = 10^8$.

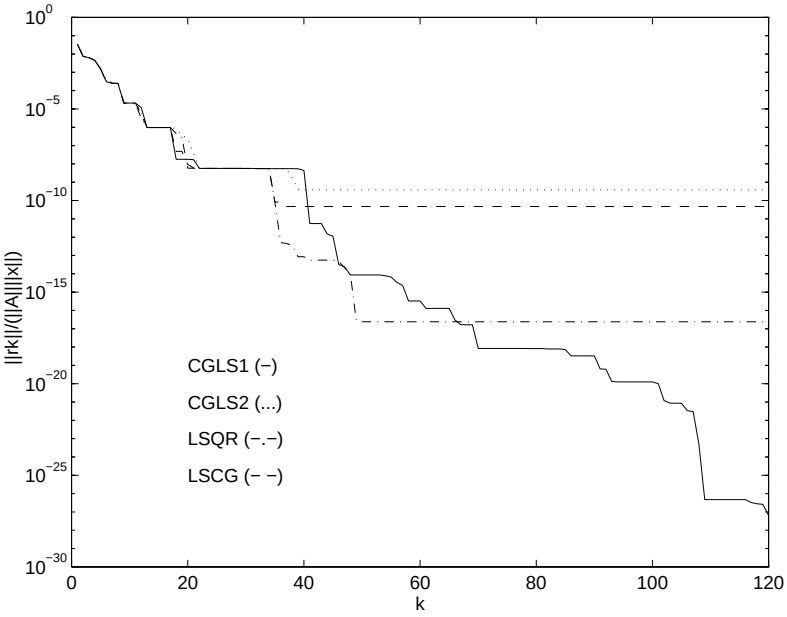


FIG. 6.2. $\|r_k\|/(\|A\|\|x\|)$ *for problem* $PS(10, 10, 1, 8)$, $\kappa = 10^8$.

The first tests were run on the consistent problem $PS(10, 10, 1, 8)$, with $\kappa(A) = 10^8$. Results are shown for CGLS1, CGLS2, and the two Lanczos methods LSQR and LSCG in Figures 6.1–6.3. Here x and r denote the exact solution and the exact residual. Hence $r = 0$ in the consistent case.

Figure 6.1 shows that CGLS1 and LSQR both achieve a relative accuracy of about 10^{-9}, whereas the error in the unstable versions CGLS2 and LSCG are about a factor

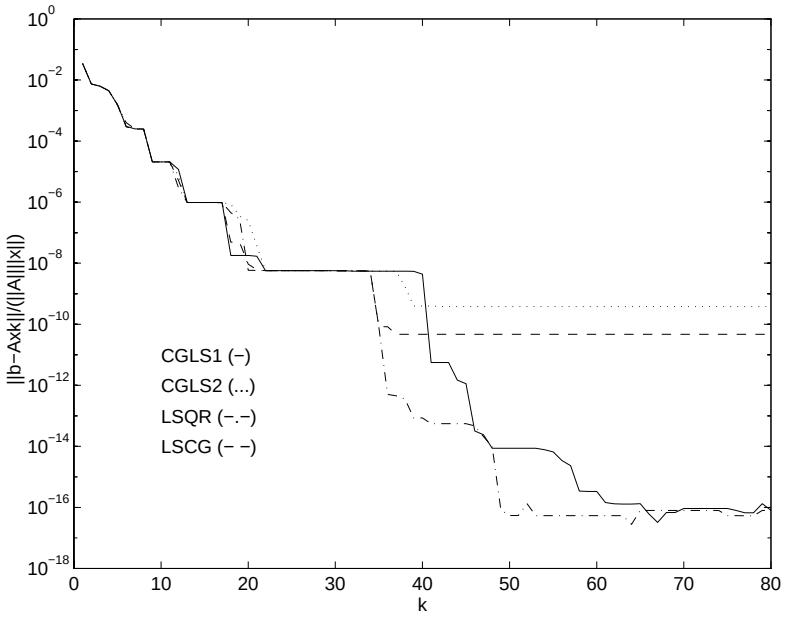


FIG. 6.3. $\|b - Ax_k\|/(\|A\|\|x\|)$ *for problem* $PS(10, 10, 1, 8)$, $\kappa = 10^8$.

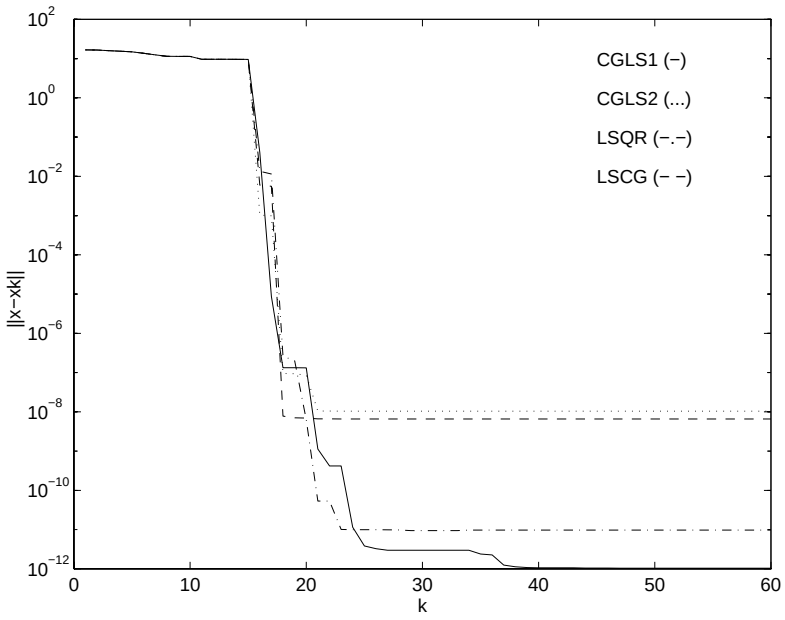


FIG. 6.4. $\|x - x_k\|$ *for problem* $PS(20, 10, 1, 4)$, $\kappa = 10^4$, $\kappa_{LS} = 6.81 \cdot 10^4$.

of $\kappa(A)$ worse. In Figure 6.2 the norm of the *recursive* residual is plotted. For CGLS1 this norm still decreased after 200 iterations, when it had reached 10^{-35}. Figure 6.3 shows the relative norm of the *true* residual, which for both stable versions reaches the level of machine precision after 50–70 iterations.

The second test problem was the inconsistent problem $PS(20, 10, 1, 4)$ with $\rho = 0.01$, for which $\kappa(A) = 10^4$, and $\kappa_{LS} = 6.81 \cdot 10^4$ (see (4.2)–(4.3)). Figure 6.4 shows

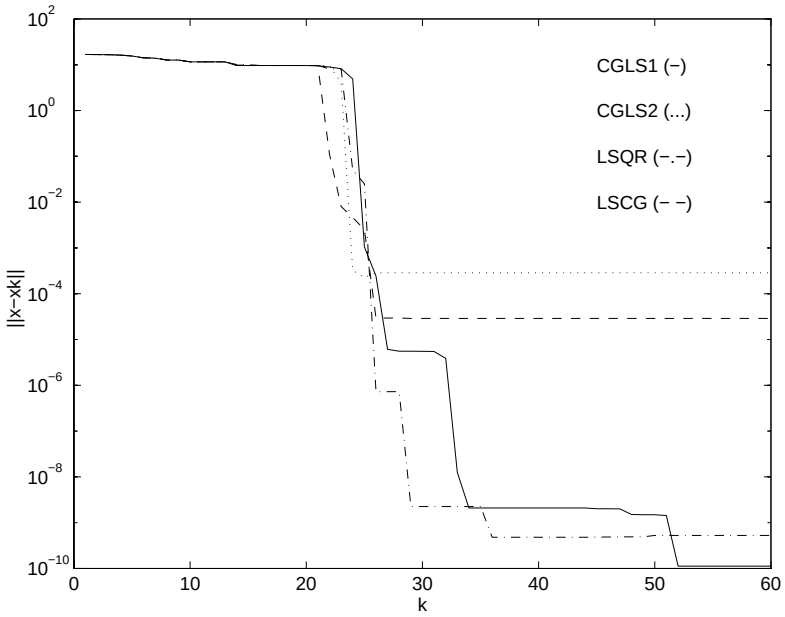


FIG. 6.5. $\|x - x_k\|$ *for problem* $PS(20, 10, 1, 6)$, $\rho = 0.001$, $\kappa = 10^6$, $\kappa_{LS} = 5.91 \cdot 10^7$.

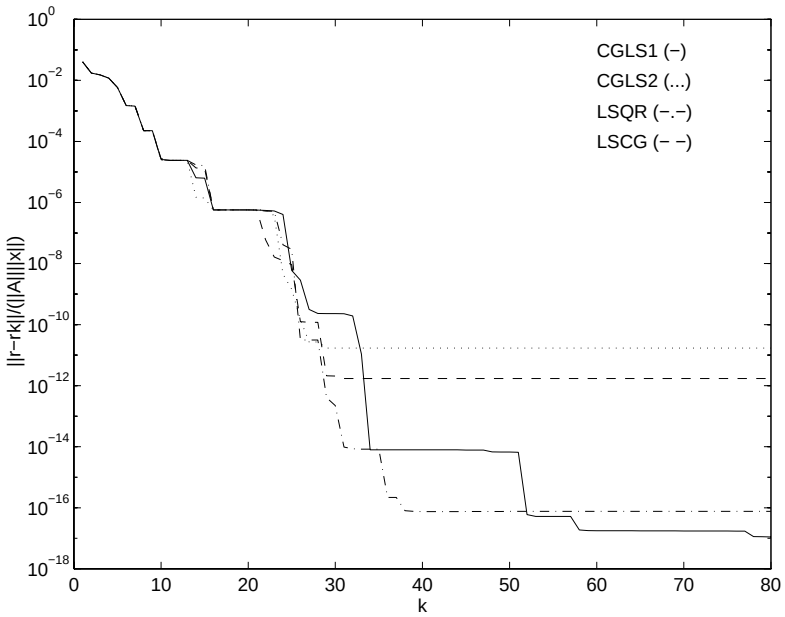


FIG. 6.6. $\|r - r_k\|/(\|A\|\|x\|)$ *for problem* $PS(20, 10, 1, 6)$, $\rho = 0.001$, $\kappa = 10^6$, $\kappa_{LS} = 5.91 \cdot 10^7$.

that the limiting relative accuracy in the solution is better than 10^{-11} for the stable versions. Again the accuracy for the unstable versions is worse by a factor of $\kappa(A)$.

In Figures 6.5–6.6 we show the results for the more ill-conditioned inconsistent problem $PS(20, 10, 1, 6)$ with $\rho = 0.001$, $\kappa(A) = 10^6$, and $\kappa_{LS} = 5.91 \cdot 10^7$. The relative accuracy in the computed solution, shown in Figure 6.5, is here around 10^{-9}

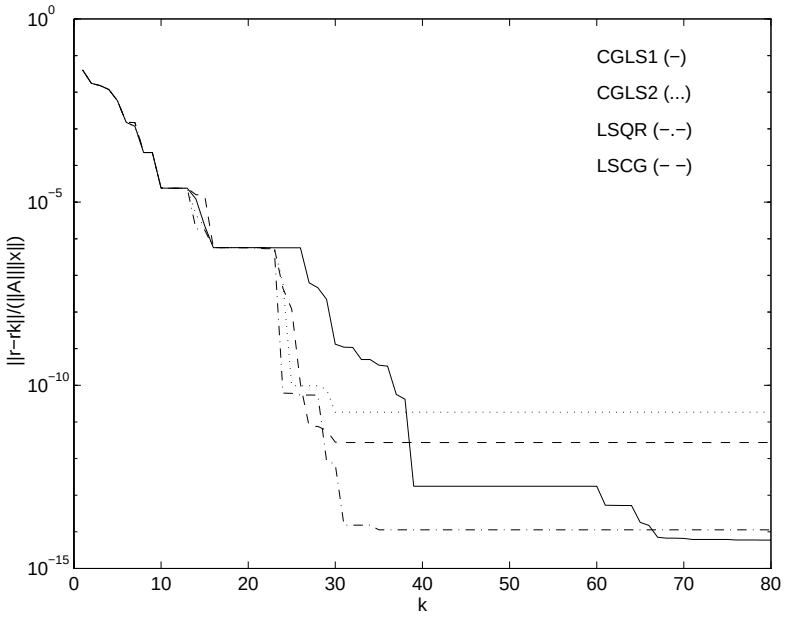


FIG. 6.7. $\|r - r_k\|/(\|A\|\|x\|)$ *for problem* $PS(20, 10, 1, 6)$, $\rho = 0.1$, $\kappa = 10^6$, $\kappa_{LS} = 5.81 \cdot 10^9$.

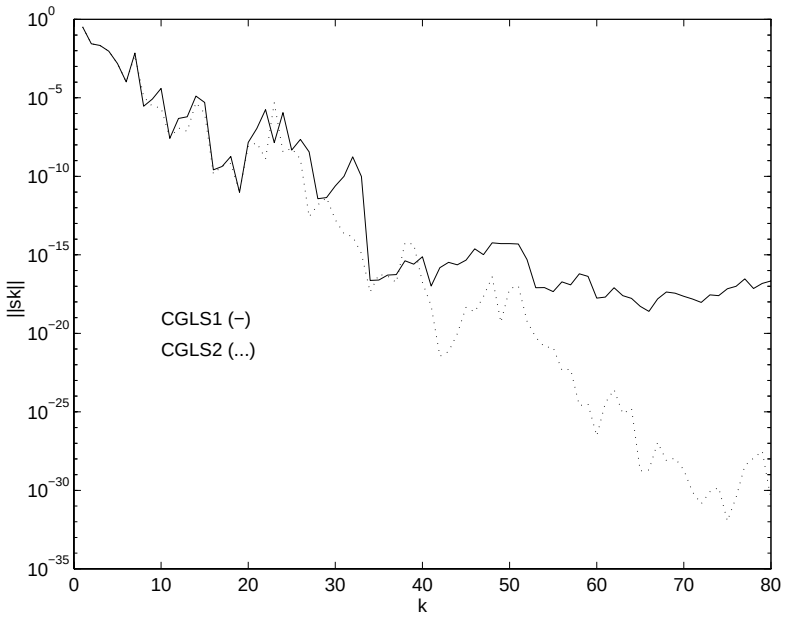


FIG. 6.8. $\|s_k\|$ *for problem* $PS(20, 10, 1, 6)$, $\rho = 0.001$, $\kappa = 10^6$, $\kappa_{LS} = 5.91 \cdot 10^7$.

for both stable versions. Figure 6.6 shows that in the limit $\|r - r_k\|/(\|A\|\,\|x\|)$ is less than 10^{-15}. In Figure 6.7 $\rho = 0.1$, $\kappa_{LS} = 5.81 \cdot 10^9$, and the limit is about 10^{-14}. Considering the ill-conditioning of the problem these results are much better than can be expected from a normwise backward stable method and support the assumption in (5.5).

Finally, in Figure 6.8 the norm of the recursive residual $s_k = A^T r_k$ of the normal

equations is plotted for the case $\rho = 0.001$. For CGLS1 this decreases to about machine precision, whereas for the unstable version CGLS2 it decreases to zero in the limit. Hence it is important *not* to base the comparison of the two algorithms on this quantity!

The results suggests some cheap stopping rules for compatible systems. As shown by Figures 6.2 and 6.3, in this case $\|r_k\|$ becomes excessively small and so will cause termination at a sensible point, even though $\|b - Ax_k\|$ will not be as small as predicted. For incompatible systems, as Figure 6.8 shows, $\|s_k\|$ levels off at $O(u)$. On the other hand, for the unstable version CGLS2, $\|s_k\|$ does eventually become very small. However, this does not mean that $\|r - (b - Ax_k)\|$ will be small and should not be used as a reason for using CGLS2.

7. Conclusions. We have studied two different implementations, CGLS1 and CGLS2, of the conjugate gradient method applied to the normal equations and two algorithms, LSQR and LSCG, based on Lanczos bidiagonalization. Although these four algorithms for solving the linear least squares problem $\min_x \|Ax - b\|$ are mathematically equivalent, their performance in finite precision differs significantly. The achievable accuracy in finite precision of CGLS2 and LSCG can be lower by a factor of

$$\frac{\|b\|}{\max\{\|r\|, \|x\|/\|A^\dagger\|\}}$$

than that of a backward stable method. For the preferred implementation CGLS1 as well as for LSQR a bound is derived for the error $\|r - r_k\|$ of the computed residual r_k. This bound shows that for the consistent case ($r = 0$) these two methods achieve an accuracy similar to a normwise backward stable method. For the inconsistent case CGLS1 and LSQR achieve even better accuracy than a normwise backward stable method. Numerical tests confirming this behavior have been given.

Acknowledgments. The authors would like to thank Tianruo Yang for carrying out most of the numerical tests. We also thank Michael Saunders for giving valuable comments on the error analysis and on stopping rules. An anonymous referee report also helped to improve the paper.

REFERENCES

[1] Å. Björck, *Conjugate Gradient Methods for Sparse Least Squares Problems*, Unpublished notes, Computer Science Department, Stanford University, Stanford, CA, 1979.
[2] Å. Björck, *Component-wise perturbation analysis and errors bounds for linear least squares solutions*, BIT, 31 (1991), pp. 238–244.
[3] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.
[4] Å. Björck, E. Grimme, and P. Van Dooren, *An implicit bidiagonalization algorithm for ill-posed systems*, BIT, 34 (1994), pp. 510–534.
[5] Y. T. Chen, *Iterative Methods for Linear Least Squares Problems*, Technical report CS-75-04, Department of Computer Science, University of Waterloo, Waterloo, ON, Canada, 1975.
[6] E. J. Craig, *The N-step iteration procedures*, J. Math. Phys., 34 (1955), pp. 64–73.
[7] T. Elfving, *On the Conjugate Gradient Method for Solving Linear Least Squares Problems*, Report LiTH-MAT-R-78-3, Department of Math., Linköping University, Linköping, Sweden, 1978.
[8] R. W. Freund, G. H. Golub, and N. Nachtigal, *Iterative solution of linear systems*, Acta Numerica, (1991), pp. 57–100.
[9] G. H. Golub and W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal. Ser. B, 2 (1965), pp. 205–224.
[10] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.

[11] A. Greenbaum, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.
[12] A. Greenbaum and Z. Strakoš, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137.
[13] M. T. Heath, *Numerical methods for large sparse linear least squares problems*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 497–513.
[14] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Stds., B49 (1952), pp. 409–436.
[15] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[16] N. J. Higham and P. A. Knight, *Finite precision behavoir of stationary iteration for solving singular systems*, Linear Algebra Appl., 192 (1993), pp. 165–186.
[17] P. Läuchli, *Iterative Lösung und Fehlerabschätzung in der Ausgleichrechnung*, ZAMP, 10 (1959), pp. 245–280.
[18] C. L. Lawson, *Sparse Matrix Methods based on Orthogonality and Conjugacy*, Technical mem. 33-627, Jet Propulsion Lab., California Institute of Technology, 1973.
[19] C. C. Paige, *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, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71.
[21] J. K. Reid, *On the method of conjugate gradients for the solution of large systems of linear equations*, in Large Sparse Sets of Linear Equations, J. K. Reid, ed., Academic Press, New York, 1971, pp. 231–254.
[22] D. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[23] E. Stiefel, *Ausgleichung ohne Aufstellung der Gausschen Normalgleichungen*, Wiss. Z. Technische Hochschule Dresden, 2 (1952/53), pp. 441–442.
[24] Z. Strakoš, *On the real convergence of the conjugate gradient method*, Linear Algebra Appl., 154/156 (1991), pp. 535–549.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 737–754, July 1998

009

PARTIAL SUPERDIAGONAL ELEMENTS AND SINGULAR VALUES OF A COMPLEX SKEW-SYMMETRIC MATRIX*

TIN-YAU TAM†

Dedicated to Professor Y. H. Au-Yeung

Abstract. Let A be an $m \times m$ complex skew-symmetric matrix with singular values $s_1 \geq s_1 \geq s_2 \geq s_2 \geq \cdots \geq s_n \geq s_n \geq 0$, where $n = [m/2]$. We consider the sets $\widetilde{D}_p(A) = \{\text{diag}\,(U^T AU[1,\ldots,p|n+1,\ldots,n+p]) : U \in U(m)\}$, $p = 1,\ldots,n$, where $U(m)$ denotes the unitary group. We prove that when $m = 2n$ and $p = n$, $d = (d_1,\ldots,d_n) \in \widetilde{D}_n(A)$ if and only if

$$\sum_{i=1}^{k} |d_i| \leq \sum_{i=1}^{k} s_i, \qquad k = 1,\ldots,n,$$
$$\sum_{i=1}^{n-1} |d_i| - |d_n| \leq \sum_{i=1}^{n-1} s_i - s_n,$$

after rearranging the entries of d in descending order with respect to absolute value. The set is not convex in general. The inequalities are identical to those of Thompson–Sing's theorem on the diagonal elements and the singular values of an $n \times n$ complex matrix.

All other cases, i.e., (1) $m = 2n+1$ and $1 \leq p \leq n$ and (2) $m = 2n$ and $1 \leq p < n$, are completely described by the inequailities $\sum_{i=1}^{k} |d_i| \leq \sum_{i=1}^{k} s_i$, $k = 1,\ldots,p$. The sets are all convex. Various applications and related results are obtained.

Key words. singular values, superdiagonal elements

AMS subject classifications. 15A18, 15A45

PII. S0895479896312559

1. Introduction.

The well-known Schur–Horn's result [3, 6] asserts that the diagonal elements of a Hermitian matrix with prescribed eigenvalues are completely described by majorization. The result is also true for the real symmetric matrices. It can be stated as

$$\{\text{diag}\,(U^T \Lambda U) : U \in SO(n)\} = \{\text{diag}\,(U^T \Lambda U) : U \in U(n)\} = \{d \in \mathbb{R}^n : d \prec \lambda\},$$

where $\Lambda = \text{diag}\,(\lambda_1,\ldots,\lambda_n)$. Here $d \prec \lambda$ means that d, a diagonal element, is majorized by λ, the eigenvalue element. We remark that $SO(n)$ can be replaced by $O(n)$, the orthogonal group and $U(n)$ can be replaced by $SU(n)$, the special unitary group, in the above expressions.

Thompson [14] and Sing [8] independently described the set of diagonal elements of matrices with prescribed singular value $s_1,\ldots,s_n$ ($\{\text{diag}\,(USV) : U, V \in U(n)\}$) in terms of inequalities. Thompson [14] also handled the cases when $U(n)$ is replaced by $SO(n)$ and $O(n)$, respectively. He [15] then obtained the characterization of the set of diagonal elements of complex symmetric matrices with prescribed singular values $\{\text{diag}\,(U^T SU) : U \in U(n)\}$.

The real skew-symmetric case has very recently been examined (Theorem 1, [10]), i.e., the description of the set $D_n(A) = \{(d_1,\ldots,d_n) : d_i = (O^T AO)_{2i-1,2i}, O \in$

* Received by the editors November 22, 1996; accepted for publication (in revised form) by R. Bhatia June 23, 1997; published electronically March 18, 1998.
http://www.siam.org/journals/simax/19-3/31255.html
† Department of Mathematics, Auburn Univeristy, AL 36849-5310 (tamtiny@mail.auburn.edu).

$SO(m)\}$ where A is a real skew-symmetric $m \times m$ matrix and $n = [m/2]$. One can easily obtain a similar result when $SO(m)$ is replaced by $O(m)$.

The set $D_n(A)$ is the collection of the superdiagonal elements $(a_{12}, a_{34}, \ldots, a_{2n-1,2n})$ of the real skew-symmetric matrices A with prescribed singular values $s_1 \geq s_1 \geq s_2 \geq s_2 \geq \cdots \geq s_n \geq s_n \geq 0$ under the conjugate action of the special orthogonal group.

The proofs of the following results are not difficult (see the proof of Theorem 6 in [14] for the second part of Corollary 1.1). The first part is indeed a corollary to Schur–Horn's result and the second part is a corollary to Theorem 1 of [10].

COROLLARY 1.1. 1. *Let $1 \leq p < n$. Then $d \in \mathbb{R}^p$ is a partial diagonal element of length p of a real symmetric matrix (Hermitian matrix) with prescribed eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ if and only if $\sum_{i=1}^k d_k \leq \sum_{i=1}^k \lambda_k$, $k = 1, \ldots, p$, after rearranging the entries of d in descending order.*

2. *Let A be an $m \times m$ real skew-symmetric matrix with singular values $s_1 \geq s_1 \geq s_2 \geq s_2 \geq \cdots \geq s_n \geq s_n \geq 0$, where $n = [m/2]$. If $1 \leq p < n$, then $d \in D_p(A)$ if and only if $\sum_{i=1}^k |d_i| \leq \sum_{i=1}^k s_i$, $k = 1, \ldots, p$, after rearranging the entries of d in descending order with respect to absolute value.*

Let $A[1, \ldots, p|n+1, \ldots, n+p]$ denote the submatrix of A lying on the rows indexed by $1, \ldots, p$ and on the columns indexed by $n+1, \ldots, n+p$. We denote the complementary submatrix by $A(1, \ldots, p|n+1, \ldots, n+p)$. We have the following reformulation.

PROPOSITION 1.2. *Let A be an $m \times m$ real skew-symmetric matrix. Let $1 \leq p \leq n$ where $n = [m/2]$.* 1. *If* (1) *$m = 2n+1$, or* (2) *$m = 2n$ and $1 \leq p < n$, or* (3) *$m = 2n$, $p = n$ and $[n/2]$ is even, then $D_p(A) = \{diag\,(U^T AU)[1, \ldots, p|n+1, \ldots, n+p] : U \in SO(m)\}$.*

2. *If $m = 2n$, and $[n/2]$ is odd, then $D_n(A) = \{diag\,(U^T(D^T AD)U)[1, \ldots, n|n+1, \ldots, 2n] : U \in SO(2n)\}$, where $D = diag\,(1, \ldots, 1, -1)$.*

Proof. Let $P \equiv P_n$ be the $2n \times 2n$ permutation matrix such that $p_{2k-1,k} = p_{2k,n+k} = 1$, $k = 1, \ldots, n$ and zero otherwise. The determinant of P is $\det P = (-1)^{n(n-1)/2} = (-1)^{[n/2]}$. By Laplace expansion and some column exchanges, we have $\det P_n = (-1)^{n-1} \det P_{n-1}$. Hence, $\det P = (-1)^{n(n-1)/2}$. Suppose that $[n/2]$ is even. (1) When n is even, $n = 4k$ and, hence, $n(n-1)/2$ is even. (2) When n is odd, $(n-1)/2 = [n/2]$ is even and, hence, $n(n-1)/2$ is even. Similarly, odd $[n/2]$ implies that $n(n-1)/2$ is odd. So $\det P = (-1)^{[n/2]}$. Moreover, for any $2n \times 2n$ matrix B,

$$\operatorname{diag} P^T BP[1, 3, \ldots, 2n-1|2, 4, \ldots, 2n] = \operatorname{diag} B[1, 2, \ldots, n|n+1, n+2, \ldots 2n].$$

When $m = 2n+1$, we set $O = P \oplus (-1)^{[n/2]} \in SO(2n+1)$. If $1 \leq p \leq n$, we have

$$\begin{aligned} D_p(A) &= \{(d_1, \ldots, d_p) : d_i = (U^T AU)_{2i-1,2i},\ i = 1, \ldots, p,\ U \in SO(2n+1)\} \\ &= \{\operatorname{diag}(U^T AU)[1, 3, \ldots, 2p-1|2, 4, \ldots, 2p] : U \in SO(2n+1)\} \\ &= \{\operatorname{diag}(O^T U^T AUO)[1, 3, \ldots, 2p-1|2, 4, \ldots, 2p] : U \in SO(2n+1)\} \\ &= \{\operatorname{diag}(U^T AU)[1, \ldots, p|n+1, \ldots, n+p] : U \in SO(2n+1)\}. \end{aligned}$$

When $m = 2n$ we have to deal with it slightly differently, since P may be in $O(2n) \backslash SO(2n)$. Let $D = \operatorname{diag}(1, \ldots, 1, -1)$. Hence,

1. If $[n/2]$ is even, i.e., $P \in SO(2n)$,

$$\begin{aligned} D_p(A) &= \{\operatorname{diag}(P^T U^T AUP)[1, 3, \ldots, 2p-1|2, 4, \ldots, 2p] : U \in SO(2n)\} \\ &= \{\operatorname{diag}(U^T AU)[1, \ldots, p|n+1, \ldots, n+p] : U \in SO(2n)\}. \end{aligned}$$

2. If $[n/2]$ is odd, i.e., $P \notin SO(2n)$,

$$\begin{aligned} D_p(A) &= \{\mathrm{diag}\,(P^TU^TD^TADUP)[1,3,\ldots,2p-1|2,4,\ldots,2p] : U \in SO(2n)\} \\ &= \{\mathrm{diag}\,(U^T(D^TAD)U)[1,\ldots,p|n+1,\ldots,n+p] : U \in SO(2n)\}. \end{aligned}$$

The above expression becomes $\{\mathrm{diag}\,(U^TAU)[1,\ldots,p|n+1,\ldots,n+p] : U \in SO(2n)\}$ if $1 \le p < n$ since

$$\begin{aligned} &\{\mathrm{diag}\,(U^TD^TADU)[1,\ldots,p|n+1,\ldots,n+p] : U \in SO(2n)\} \\ &= \{\mathrm{diag}\,(D^TU^TAUD)[1,\ldots,p|n+1,\ldots,n+p] : U \in SO(2n)\} \\ &= \{\mathrm{diag}\,(U^TAU)[1,\ldots,p|n+1,\ldots,n+p] : U \in SO(2n)\}. \qquad \square \end{aligned}$$

When $m = 2n$ and if the canonical form of the real skew-symmetric matrix A is

$$\begin{pmatrix} 0 & s_1 \\ -s_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & s_{n-1} \\ -s_{n-1} & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & s'_n \\ -s'_n & 0 \end{pmatrix},$$

then

$$\begin{pmatrix} 0 & s_1 \\ -s_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & s_{n-1} \\ -s_{n-1} & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & -s'_n \\ s'_n & 0 \end{pmatrix}$$

is the canonical form of D^TAD. Here $s'_n = \pm s_n$.

It is then natural to ask the complex skew-symmetric case. If A is an $m \times m$ complex skew-symmetric matrices, there exists $U \in U(m)$ such that

$$U^TAU = \begin{cases} \begin{pmatrix} 0 & s_1 \\ -s_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & s_n \\ -s_n & 0 \end{pmatrix} \oplus 0 & \text{if } m = 2n+1, \\ \begin{pmatrix} 0 & s_1 \\ -s_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & s_n \\ -s_n & 0 \end{pmatrix} & \text{if } m = 2n. \end{cases}$$

Clearly, the numbers $s_1 \ge s_1 \ge s_2 \ge s_2 \ge \cdots \ge s_n \ge s_n \ge 0$ are the singular values of A. Since the underlying group is $U(m)$, the decomposition is simpler than that of the real skew-symmetric case under the conjugation of $SO(m)$ in which we have two possibilities with respect to the case $m = 2n$.

Our first objective is to study the set $\widetilde{D}_p(A) = \{\mathrm{diag}\,(U^TAU)[1,3,\ldots,2p-1|2,4,\ldots,2p] : U \in U(m)\}$, where $1 \le p \le n$, $n = [m/2]$, and A is an $m \times m$ complex skew-symmetric matrix. According to the discussion of the real analogy (see the proof of Proposition 1.2),

$$\begin{aligned} \widetilde{D}_p(A) &= \{\mathrm{diag}\,(U^TAU)[1,\ldots,p|n+1,\ldots,n+p] : U \in U(m)\} \\ &= \{\mathrm{diag}\,(U^TS_{n,n}U)[1,\ldots,p|n+1,\ldots,n+p] : U \in U(m)\}, \end{aligned}$$

where $S = \mathrm{diag}\,(s_1,\ldots,s_n)$ and

$$S_{n,n} = \begin{pmatrix} 0 & S \\ -S & 0 \end{pmatrix}.$$

While comparing Schur–Horn's result (real symmetric case) and Thompson's result (Theorem 1, [15], complex symmetric case), it seems that the real skew-symmetric case (Theorem 1 in [10]) and the complex skew-symmetric case behave quite differently. Surprisingly, this is not so as we see Theorem 2.1 and Theorem 2.2 in the

next section. In section 3, we have some applications of Theorem 1 of [10], Theorem 2.1, and Theorem 2.2, mainly concerning the singular values inequalities of some submatrix of a real or complex skew-symmetric matrix. In section 4, we have some discussion and make further applications. In the final section, we provide a list of related results.

2. Partial superdiagonal elements and singular values. The following two theorems are the main results in this section. Theorem 2.1 is for the even case and Theorem 2.2 is for the odd case.

THEOREM 2.1. *Let A be a $2n \times 2n$ complex skew-symmetric matrix with singular values $s_1 \geq s_1 \geq s_2 \geq s_2 \geq \cdots \geq s_n \geq s_n$.*

1. *Then $d \in \widetilde{D}_n(A)$ if and only if*

$$\sum_{i=1}^{k} |d_i| \leq \sum_{i=1}^{k} s_i, \qquad k = 1, \ldots, n, \tag{1}$$

$$\sum_{i=1}^{n-1} |d_i| - |d_n| \leq \sum_{i=1}^{n-1} s_i - s_n, \tag{2}$$

after rearranging the entries of d in descending order with respect to modulus. Hence $\widetilde{D}_n(A) \subset \mathbb{C}^n$ is not convex in general.

2. *When $1 \leq p < n$, $d \in \widetilde{D}_p(A)$ if and only if*

$$\sum_{i=1}^{k} |d_i| \leq \sum_{i=1}^{k} s_i, \qquad k = 1, \ldots, p, \tag{3}$$

after rearranging the entries of d in descending order with respect to modulus. Hence $\widetilde{D}_p(A) \subset \mathbb{C}^p$ is a convex set.

While $s_n = 0$, the set $\widetilde{D}_n(A)$ has been studied [9]. When $s_n = 0$, 2 becomes superficial and the set $\widetilde{D}_n(A)$ is then convex.

The inequalities 1 and 2 are identical to those of Thompson [14] and Sing [8]. See [11] for the Lie explanation of the result of Thompson and Sing.

THEOREM 2.2. *Let A be a $(2n+1) \times (2n+1)$ complex skew-symmetric matrix with singular values $s_1 \geq s_1 \geq s_2 \geq s_2 \geq \cdots \geq s_n \geq s_n \geq 0$. For $1 \leq p \leq n$, the vector $d \in \mathbb{C}^p$ is an element of $\widetilde{D}_p(A)$ if and only if* 3 *is satisfied. Hence, $\widetilde{D}_p(A) \subset \mathbb{C}^p$ is a convex set.*

The cases $1 \leq p < n$ and the special case $p = n$ and $s_n = 0$ of Theorem 2.2 were obtained in [9].

We first observe that if A is an $m \times m$ complex skew-symmetric matrix, then $\mathcal{T}(\widetilde{D}_p(A)) = \widetilde{D}_p(A)$. Here $\mathcal{T}(K)$ is the torus generated by $K \subset \mathbb{C}^p$, i.e.,

$$\mathcal{T}(K) = \{(e^{i\theta_1} d_1, \ldots, e^{i\theta_p} d_p) : (d_1, \ldots, d_p) \in K,\ \theta_1, \ldots, \theta_p \in \mathbb{R}\}.$$

Of course, this is also true for the complex symmetric case. It is because if $d \in \widetilde{D}_p(A)$, i.e., $d = \operatorname{diag}(U^T A U)[1, \ldots, p | n+1, \ldots, n+p]$ for some $U \in U(m)$ where $n = [m/2]$, then $(e^{i\theta_1} d_1, \ldots, e^{i\theta_p} d_p) = \operatorname{diag}(V^T U^T A U V)[1, \ldots, p | n+1, \ldots, n+p]$ where

$$V = \operatorname{diag}(e^{i\theta_1/2}, \ldots, e^{i\theta_p/2}) \oplus I_{n-p} \oplus \operatorname{diag}(e^{i\theta_1/2}, \ldots, e^{i\theta_p/2}) \oplus I_{n-p}(\oplus 1) \in U(m).$$

Moreover, $d \in \widetilde{D}_p(A)$ implies that $d_\sigma \equiv (d_{\sigma(1)}, \ldots, d_{\sigma(p)}) \in \widetilde{D}_p(A)$ for any $\sigma \in \Sigma_p$ (the symmetric group). It is because that if $d \in \widetilde{D}_p(A)$, i.e., $d = \operatorname{diag}(U^T A U)[1, \ldots, p|n+1, \ldots, n+p]$ for some $U \in U(m)$, then $(d_{\sigma(1)}, \ldots, d_{\sigma(p)}) = \operatorname{diag}(Q^T U^T A U Q)[1, \ldots, p|n+1, \ldots, n+p]$ where $Q = P_\sigma \oplus I_{n-p} \oplus P_\sigma \oplus I_{n-p}(\oplus 1)$ and P_σ is the permutation matrix corresponding to σ.

We will make use of these two observations freely. We also use $|d| \prec_w s$ to denote the relation 1.

LEMMA 2.3. *Let*

$$A = \begin{pmatrix} 0 & f & a & 0 \\ -f & 0 & 0 & d \\ -a & 0 & 0 & g \\ 0 & -d & -g & 0 \end{pmatrix} \in \mathbb{C}_{4\times 4}$$

be a complex skew-symmetric matrix having singular values $s_1 \geq s_2 \geq s_2 \geq s_2$ such that $a \geq 0 \geq d$, $a \geq |d|$. Then $s_1 - s_2 \geq a - |d|$ with equality holds if and only if (1) *when $ad \neq 0$, $\overline{f} = g$,* (2) *when $ad = 0$, $|f| = |g|$.*

Proof. We consider the square of Frobenius norm of A, i.e., $\operatorname{tr} A^* A$ which is also the sum of the square of the singular values of A. So we have $s_1^2 + s_2^2 = |a|^2 + |d|^2 + |f|^2 + |g|^2$. Subtracting $2s_1 s_2$ from both sides, we have $(s_1 - s_2)^2 = |a|^2 + |d|^2 + |f|^2 + |g|^2 - 2s_1 s_2$. But $s_1^2 s_2^2 = |\det A|$. By direct computation, $\det A = (ad - fg)^2$. Hence, $s_1 s_2 = |ad - fg|$. As a result, we have

$$\begin{aligned} (s_1 - s_2)^2 &= |a|^2 + |d|^2 + |f|^2 + |g|^2 - 2|ad - fg| \\ &\geq |a|^2 + |d|^2 + |f|^2 + |g|^2 - 2|ad| - 2|fg| \\ &= (|a| - |d|)^2 + (|f| - |g|)^2 \\ &\geq (a - |d|)^2. \end{aligned}$$

Taking square root yields $s_1 - s_2 \geq a - |d|$. Equality holds if and only if ad and $-fg$ are on the same ray through the origin and $|f| = |g|$. So (1) if $ad \neq 0$, i.e., $ad < 0 (\in \mathbb{R})$, then $g = \overline{f}$; (2) if $ad = 0$, then $|f| = |g|$. ∎

LEMMA 2.4. *Let*

$$A = \begin{pmatrix} 0 & f & a & 0 \\ -f & 0 & 0 & d \\ -a & 0 & 0 & g \\ 0 & -d & -g & 0 \end{pmatrix} \in \mathbb{C}_{4\times 4}$$

be a complex skew-symmetric matrix having singular values $s_1 \geq s_2 \geq s_2 \geq s_2$ such that $a, d \geq 0$. Then $s_1 + s_2 \geq a + d$ with equality holds if and only if $ad - fg \geq 0$ and $\overline{f} = g$.

Proof. Similar to the proof of Lemma 2.3, we have $(s_1 + s_2)^2 = |a|^2 + |d|^2 + |f|^2 + |g|^2 + 2s_1 s_2$. As a result, we have

$$\begin{aligned} (s_1 + s_2)^2 &= |a|^2 + |d|^2 + |f|^2 + |g|^2 + 2|ad - fg| \\ &\geq |a|^2 + +|d|^2 + |f|^2 + |g|^2 + 2|ad| - 2|fg| \\ &= (|a| + |d|)^2 + (|f| - |g|)^2 \\ &\geq (|a| + |d|)^2. \end{aligned}$$

Taking square root yields $s_1 + s_2 \geq a + d$ as $a, d \geq 0$. Equality holds if and only if $|f| = |g|$ and equality occurs in the above triangle inequality, i.e., ad and fg are on

the same half-line through the origin and $ad - fg \geq 0$. In other words, $f = \overline{g}$ and $ad - fg \geq 0$. □

Let $\mathcal{W}_{\mathbb{C}} \subset \mathbb{C}_{n\times n}$ ($\mathcal{W}_{\mathbb{R}} \subset \mathbb{R}_{n\times n}$) be the set of $n \times n$ absolutely complex (real) doubly substochastic matrices, i.e., $W \in \mathcal{W}_{\mathbb{C}}$ ($\mathcal{W}_{\mathbb{R}}$) if $\sum_{i=1}^n |w_{ij}| = \sum_{j=1}^n |w_{ij}| \leq 1$ for all $i, j = 1, \ldots, n$.

The set $\mathcal{W}_{\mathbb{C}}s$ can be interpreted as the torus $\{(e^{i\theta_1}d_1, \ldots, e^{i\theta_n}d_n) : d \in \mathbb{R}^n,$ $|d| \prec_w s\}$ generated by the set $\{d \in \mathbb{R}^n : |d| \prec_w s\} = \hat{G}(n)s$ which is identical to $\mathcal{W}_{\mathbb{R}}s$ [11].

LEMMA 2.5. *Let $Z \in U(m)$ and $n = [m/2]$ such that*

$$Z[1, \ldots, 2n|1, \ldots, 2n] = \begin{pmatrix} U & V \\ W & X \end{pmatrix}.$$

Then the matrix $(U \circ X - V \circ W)$ belongs to $\mathcal{W}_{\mathbb{C}}$ where $A \circ B$ denotes the Hadamard product of A and B.

Proof. Notice that

$$\begin{aligned}
\sum_{i=1}^n |u_{ij}x_{ij} - v_{ij}w_{ij}| &\leq \sum_{i=1}^n |u_{ij}x_{ij}| + \sum_{i=1}^n |v_{ij}w_{ij}| \\
&\leq \left(\sum_{i=1}^n |u_{ij}|^2 \sum_{i=1}^n |x_{ij}|^2 \right)^{1/2} + \left(\sum_{i=1}^n |v_{ij}|^2 \sum_{i=1}^n |w_{ij}|^2 \right)^{1/2} \\
&\leq \left(\sum_{i=1}^n |u_{ij}|^2 \sum_{i=1}^n |x_{ij}|^2 \right)^{1/2} \\
&\quad + \left[\left(1 - \sum_{i=1}^n |x_{ij}|^2 \right) \left(1 - \sum_{i=1}^n |u_{ij}|^2 \right) \right]^{1/2} \\
&\leq \frac{1}{2} \left(\sum_{i=1}^n |u_{ij}|^2 + \sum_{i=1}^n |x_{ij}|^2 \right) + \frac{1}{2} \left[\left(1 - \sum_{i=1}^n |x_{ij}|^2 \right) \right. \\
&\quad \left. + \left(1 - \sum_{i=1}^n |u_{ij}|^2 \right) \right] \\
&\leq 1.
\end{aligned}$$

The first inequality is simply triangle inequality. The second is Cauchy–Schwarz inequality. The second last inequality is due to arithmetic-geometric mean inequality. The third inequality is due to the fact that Z is a unitary matrix. Similarly, $\sum_{j=1}^n |u_{ij}x_{ij} - v_{ij}w_{ij}| \leq 1$. In other words, the matrix $U \circ X - V \circ W$ is an $n \times n$ absolutely doubly substochastic matrix. □

The following result is due to Thompson (Theorem 5 in [14]).

LEMMA 2.6. [14]. *Let $s = (s_1, \ldots, s_n) \in \mathbb{R}^n_+$. Then $\mathcal{W}_{\mathbb{C}}s = \{d \in \mathbb{C}^n : |d| \prec_w s\}$ and $\mathcal{W}_{\mathbb{R}}s = \{d \in \mathbb{R}^n : |d| \prec_w s\}$.*

LEMMA 2.7. *Let A be an $m \times m$ complex skew-symmetric matrix. If $d \in \widetilde{D}_n(A)$, then $|d| \prec_w s$.*

Proof. Let $Z \in U(m)$ and $n = [m/2]$ such that $d = \operatorname{diag}(ZS_{n,n}Z^T)[1, \ldots, n|n + 1, \ldots, 2n]$. Let

$$Z[1, \ldots, 2n|1, \ldots, 2n] = \begin{pmatrix} U & V \\ W & X \end{pmatrix}.$$

Then $d = \operatorname{diag}(USX^T - VSW^T)$ by direct computation and hence $d = (U \circ X - V \circ W)s$ where $A \circ B$ denotes the Hadamard product of A and B. Hence, by Lemma 2.5 and 2.6, the result follows. □

Remark. One can apply the necessity part of Thompson–Sing's result. Our approach here reveals more than the statement of Lemma 2.7, e.g., Lemma 2.5 shows that the matrix $U \circ X - V \circ W$ is absolutely substochastic. It is a more fundamental reason why Lemma 2.7 is true.

LEMMA 2.8. *Let A be a $2n \times 2n$ complex skew-symmetric matrix with singular values $s_1 \geq s_1 \geq s_2 \geq s_2 \geq \cdots \geq s_n \geq s_n$. If $d \in \widetilde{D}_n(A)$, then (2) is satisfied.*

Proof. Like the approach of Thompson [14] and Sing [8], we consider the following function φ. Indeed the idea can be traced back to Horn's paper [3], which was originated from von Neumann ([3], p. 628).

Define a map $\varphi : \{U^T S_{n,n} U : U \in U(2n)\} \to \mathbb{R}$ by $\varphi(U^T S_{n,n} U) = |a_{11}| + \cdots + |a_{n-1,n-1}| - |a_{nn}|$, where $A = U^T S_{n,n} U[1, \ldots, n | n+1, \ldots, 2n]$. It is clearly continuous on the compact set $\{U^T S_{n,n} U : U \in U(2n)\}$. So M, the maximum value of φ, exists. Obviously, $M = \varphi(S_{n,n}) \geq s_1 + \cdots + s_{n-1} - s_n$. We wish to show that $M = s_1 + \cdots + s_{n-1} - s_n$. Matrices $U^T S_{n,n} U$ are called maximal matrices if $M = \varphi(U^T S_{n,n} U)$.

If $n = 1$, it is trivial because $U^T S_{1,1} U$ is always

$$\begin{pmatrix} 0 & s_1 \\ -s_1 & 0 \end{pmatrix}, \qquad \text{or} \qquad \begin{pmatrix} 0 & -s_1 \\ s_1 & 0 \end{pmatrix},$$

for any $U \in U(2)$, i.e., φ is a constant function.

Suppose $n \geq 2$. Let $U^T S_{n,n} U = \begin{pmatrix} F & A \\ -A^T & G \end{pmatrix}$ be a maximal matrix and let $U^T A V = \Sigma \equiv \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, -\sigma_n)$, $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$, where U and V are $n \times n$ unitary matrices. Then consider

$$\begin{pmatrix} U^T F U & U^T A V \\ -V^T A^T U & V^T G V \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}^T \begin{pmatrix} F & A \\ -A^T & G \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}.$$

By Thompson–Sing's result, $|a_{11}| + \cdots + |a_{n-1,n-1}| - |a_{nn}| \leq \sigma_1 + \cdots + \sigma_{n-1} - \sigma_n$. So we have a maximal matrix of the form

$$B = \begin{pmatrix} F & \Sigma \\ -\Sigma & G \end{pmatrix},$$

where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, -\sigma_n)$, $F^T = -F$, and $G^T = -G$. We can assume that $\sigma_{n-1} > 0$ for if $\sigma_{n-1} = 0$, then $\sigma_1 + \cdots + \sigma_{n-1} - \sigma_n = \sigma_1 + \cdots + \sigma_{n-2} \leq s_1 + \ldots + s_{n-2} + s_{n-1} - s_n$. Furthermore, we assume that $s_n > 0$. Otherwise, $s_1 + \cdots + s_{n-1} - s_n = s_1 + \cdots + s_{n-1} \geq \sigma_1 + \cdots + \sigma_{n-1} \geq \sigma_1 + \cdots + \sigma_{n-1} - \sigma_n$.

We consider two cases: (a) $\sigma_n > 0$; (b) $\sigma_n = 0$.

Suppose that (a) happens. Then we claim that $\overline{f}_{pn} = g_{pn}$ for all $1 \leq p < n$. For if $\overline{f}_{pn} \neq g_{pn}$, then we apply Lemma 2.3 on the submatrix

$$B[p, n, n+p, 2n | p, n, n+p, 2n] = \begin{pmatrix} 0 & f_{pn} & \sigma_p & 0 \\ -f_{pn} & 0 & 0 & -\sigma_n \\ -\sigma_p & 0 & 0 & g_{pn} \\ 0 & \sigma_n & -g_{pn} & 0 \end{pmatrix}.$$

There exists a 4×4 unitary matrix

$$Z = \begin{pmatrix} U & V \\ W & X \end{pmatrix},$$

such that

$$\begin{aligned} & Z^T B[p, n, n+p, 2n|p, n, n+p, 2n] Z \\ &= \begin{pmatrix} U^T & W^T \\ V^T & X^T \end{pmatrix} B[p, n, n+p, 2n|p, n, n+p, 2n] \begin{pmatrix} U & V \\ W & X \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & \sigma'_p & 0 \\ 0 & 0 & 0 & -\sigma'_n \\ -\sigma'_p & 0 & 0 & 0 \\ 0 & \sigma'_n & 0 & 0 \end{pmatrix}, \end{aligned}$$

where $\sigma'_p - \sigma'_n > \sigma_p - \sigma_n$. Let Z' be the $2n \times 2n$ unitary matrix such that $Z'[p, n, n+p, 2n|p, n, n+p, 2n] = Z$ and $Z'(p, n, n+p, 2n|p, n, n+p, 2n) = I_{2n-4}$. Then $B' \equiv Z'^T B Z' = \begin{pmatrix} F' & \Sigma' \\ -\Sigma' & G' \end{pmatrix}$ where $\Sigma' = \operatorname{diag}(\sigma_1, \ldots, \sigma_{p-1}, \sigma'_p, \sigma_{p+1}, \ldots, \sigma_{n-1}, -\sigma'_n)$, $\sigma'_n > 0$. So $\sigma_1 + \cdots + \sigma_{n-1} - \sigma_n < \sigma_1 + \cdots + \sigma_{p-1} + \sigma'_p + \sigma_{p+1} + \cdots + \sigma_{n-1} - \sigma'_n$, contradicting the maximal property of σ.

Suppose that (b) happens. If $f_{pn} = g_{pn} = 0$ for all $1 \leq p < n$, then the nth and the last rows and columns of B are zero vectors. In this case, $s_n = 0$. So it is sufficient to consider the case that f_{pn} or g_{pn} is not zero for some $1 \leq p < n$. We consider the submatrix

$$B[p, n, n+p, 2n|p, n, n+p, 2n] = \begin{pmatrix} 0 & f_{pn} & \sigma_p & 0 \\ -f_{pn} & 0 & 0 & 0 \\ -\sigma_p & 0 & 0 & g_{pn} \\ 0 & 0 & -g_{pn} & 0 \end{pmatrix},$$

when $1 \leq p < n$. There exists a 4×4 unitary matrix

$$Z = \begin{pmatrix} U & V \\ W & X \end{pmatrix},$$

such that

$$\begin{aligned} & Z^T B[p, n, n+p, 2n|p, n, n+p, 2n] Z \\ &= \begin{pmatrix} U^T & W^T \\ V^T & X^T \end{pmatrix} B[p, n, n+p, 2n|p, n, n+p, 2n] \begin{pmatrix} U & V \\ W & X \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & \sigma'_p & 0 \\ 0 & 0 & 0 & -\sigma'_n \\ -\sigma'_p & 0 & 0 & 0 \\ 0 & \sigma'_n & 0 & 0 \end{pmatrix}, \end{aligned}$$

where $\sigma'_p \geq \sigma'_p \geq \sigma'_n \geq \sigma'_n$ are the singular values of $B[p, n, n+p, 2n|p, n, n+p, 2n]$. By considering the first and third row of the matrix $B[p, n, n+p, 2n|p, n, n+p, 2n]$, interlacing inequalities for singular values of submatrices imply that

$$|f_{pn}|^2 + \sigma_p^2 = \sigma_p'^2, \quad |g_{pn}|^2 + \sigma_p^2 = \sigma_p'^2.$$

So if either $|f_{pn}|$ or $|g_{pn}|$ is not zero, then $\sigma'_p > \sigma_p$. Since B is a maximal matrix by Lemma 2.3, $\sigma'_p - \sigma'_n = \sigma_p$. Hence, $\sigma'_n = \sigma'_p - \sigma_p > 0$. Let Z' be the $2n \times 2n$ unitary matrix such that $Z'[p, n, n+p, 2n|p, n, n+p, 2n] = Z$ and $Z'(p, n, n+p, 2n|p, n, n+p, 2n) = I_{2n-4}$. Then $B' \equiv Z'^T B Z' = \begin{pmatrix} F' & \Sigma' \\ -\Sigma' & G' \end{pmatrix}$ where $\Sigma' = \operatorname{diag}(\sigma_1, \ldots, \sigma_{p-1}, \sigma'_p, \sigma_{p+1}, \ldots, \sigma_{n-1}, -\sigma'_n)$, $\sigma'_n > 0$ and $\sigma_{n-1} > 0$. So by case (a), we conclude that $\overline{f}'_{pn} = g'_{pn}$ where $1 \le p < n$. Notice that $\sigma_{n-1} \ge \sigma'_n$; otherwise, $\sigma_1 + \cdots + \sigma_{n-1} - \sigma_n = \sigma_1 + \cdots + \sigma_{p-1} + \sigma'_p + \sigma_{p+1} + \cdots + \sigma_{n-1} - \sigma'_n < \sigma_1 + \cdots + \sigma_{p-1} + \sigma'_p + \sigma_{p+1} + \cdots + \sigma_{n-2} + \sigma'_n - \sigma_{n-1}$ contradicting the fact that B is maximal.

In both cases, we have a maximal matrix B of the form

$$B = \begin{pmatrix} F & \Sigma \\ -\Sigma & G \end{pmatrix},$$

with $\overline{f}_{pn} = g_{pn}$, $1 \le p < n$, and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, -\sigma_n)$ with $\sigma_n > 0$.

Let us consider the submatrix $B[p, q, n+p, n+q|p, q, n+p, n+q]$, for $1 \le p < q \le n-1$. By Lemma 2.4, we can deduce that $\overline{f}_{pq} = g_{pq}$. Hence, we have a maximal matrix of the form

$$B = \begin{pmatrix} F & \Sigma \\ -\Sigma & \overline{F} \end{pmatrix},$$

with $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_{n-1}, -\sigma_n)$, $\sigma_1 \ge \cdots \ge \sigma_n > 0$. The Hermitian matrix

$$\hat{B} = \begin{pmatrix} \Sigma & F \\ -\overline{F} & \Sigma \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix} B \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}$$

also has singular values $s_1, s_1, s_2, s_2, \ldots, s_n, s_n$ which are the absolute values of the eigenvalues of $\hat{B}$. Moreover, the eigenvalues of $\hat{B}$ have even algebraic multiplicities, i.e., the eigenvalues of $\hat{B}$ are $\lambda_1, \lambda_1, \lambda_2, \lambda_2, \ldots, \lambda_n, \lambda_n$. (One way to see it is by considering $\mathfrak{gl}_n(\mathbb{H})$ where $\mathbb{H}$ denotes the skew field of quaternions. A translation is needed; see p. 310 in [7].) Hence $\lambda_1, \ldots, \lambda_n$ are among $\pm s_1, \ldots, \pm s_n$. The matrix $\hat{B}$ is indefinite since $-\sigma_n < 0$ is one of the diagonal elements of $\hat{B}$. Hence, the trace of $\hat{B}$ is at most $2(s_1 + \ldots + s_{n-1} - s_n)$. Now $2(\sigma_1 + \ldots + \sigma_{n-1} - \sigma_n)$ is the trace of $\hat{B}$ and the proof is complete. □

Remark. One may try to apply Thompson–Sing's result to show the necessity of 2 in Theorems 2.1 and 2.2. But it only implies that $|d_1| + \cdots + |d_{n-1}| \le s_1 + \cdots + s_{n-1}$ instead.

Proof of Theorem 3. In view of Lemmas 2.7 and 2.8, it is adequate to show the sufficiency part.

Case 1. $k = n$. Suppose that d and s satisfy 1 and 2, then there exist unitary U and V such that $\operatorname{diag}(U^T S V) = d$, by Thompson–Sing's result. Then consider

$$\begin{pmatrix} 0 & U^T S V \\ -V^T S U & 0 \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}^T \begin{pmatrix} 0 & S \\ -S & 0 \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}.$$

An alternative is to employ Theorem 1 of [10]. For if d and s satisfy 1 and 2, the vector $|d|$ is an element of $D_n(S_{n,n})$. Then a suitable rotation makes $d \in \widetilde{D}_n(S_{n,n})$. Such an approach discloses that $SO(2n)$ and the rotations $(e^{i\theta_1}, \ldots, e^{i\theta_n})$ generate the whole $\widetilde{D}_n(A)$.

Case 2. $1 \le p < n$. The set $\widetilde{D}_p(A)$ is just the projection of $\widetilde{D}_n(A)$ through the projection $d = (d_1, \ldots, d_n) \mapsto (d_1, \ldots, d_p)$. See Theorem 6 in [14]. □

Proof of Theorem 4. Similarly, it is adequate to prove the sufficiency part.

Case 1. $k = n$. Suppose that $|d| \prec_w s$, then by Theorem 1 of [10], there exist $Z \in SO(2n+1)$ such that $|d| = \operatorname{diag}(Z^T S_{n,n} Z)$. The result follows as we are free to rotate the entries of d.

Case 2. Like the even case. □

COROLLARY 2.9. *Let* $S = diag(s_1, \ldots, s_n)$ *where* $s_1 \ge \cdots \ge s_n \ge 0$.

1. *The vector* $d = (d_1, \ldots, d_n) \in \mathbb{C}^n$ *is an element of the set* $\{diag(USX^T - VSW^T) : \begin{pmatrix} U & V \\ W & X \end{pmatrix} \in U(2n)\}$ *if and only if* (1) *and* (2) *hold.*

2. *The vector* $d = (d_1, \ldots, d_n) \in \mathbb{C}^n$ *is an element of the set* $\{diag(USX^T - VSW^T) : Z[1, \ldots, 2n|1, \ldots, 2n] = \begin{pmatrix} U & V \\ W & X \end{pmatrix}, Z \in U(2n+1)\}$ *if and only if* $|d| \prec_w s$.

Proof. Direct computation shows that

$$\widetilde{D}_n(S_{n,n}) = \begin{cases} \{\operatorname{diag}(USX^T - VSW^T) : \begin{pmatrix} U & V \\ W & X \end{pmatrix} \in U(2n)\} & \text{if } m = 2n, \\ \{\operatorname{diag}(USX^T - VSW^T) : Z[1, \ldots, 2n|1, \ldots, 2n] \\ = \begin{pmatrix} U & V \\ W & X \end{pmatrix},\ Z \in U(2n+1)\} & \text{if } m = 2n+1. \end{cases}$$ □

COROLLARY 2.10. *Let* $s = (s_1, \ldots, s_n)^T$ *where* $s_1 \ge \cdots \ge s_n \ge 0$. *Let* $A \circ B$ *denote the Hadamard product of* A *and* B.

1. *The vector* $d = (d_1, \ldots, d_n)^T \in \mathbb{C}^n$ *is an element of the set* $\{(U \circ X - V \circ W)s : \begin{pmatrix} U & V \\ W & X \end{pmatrix} \in U(2n)\}$ *if and only if* (1) *and* (2) *hold.*

2. *The vector* $d = (d_1, \ldots, d_n)^T \in \mathbb{C}^n$ *is an element of the set* $\{(U \circ X - V \circ W)s : Z[1, \ldots, 2n|1, \ldots, 2n] = \begin{pmatrix} U & V \\ W & X \end{pmatrix}, Z \in U(2n+1)\}$ *if and only if* $|d| \prec_w s$.

Proof. Direct computation shows that

$$\widetilde{D}_n(S_{n,n}) = \begin{cases} \{(U \circ X - V \circ W)s : \begin{pmatrix} U & V \\ W & X \end{pmatrix} \in U(2n)\} & \text{if } m = 2n, \\ \{(U \circ X - V \circ W)s : Z[1, \ldots, 2n|1, \ldots, 2n] \\ = \begin{pmatrix} U & V \\ W & X \end{pmatrix},\ Z \in U(2n+1)\} & \text{if } m = 2n+1. \end{cases}$$ □

Though convexity may fail as we see the complex symmetric case (Theorem 1 in [16]), Thompson–Sing's result, and the complex skew-symmetric case (Theorem 2.1), if we consider the real parts of those sets, we have convexity. (Imaginary parts work, too. Indeed, one can consider the projection of the sets onto $e^{i\theta}\mathbb{R}^n$ where $\theta \in \mathbb{R}$ is fixed and the results are valid.) See [11] and the proof of the following is omitted.

THEOREM 2.11. *Let* A *be an* $m \times m$ *complex skew-symmetric matrix with singular values* $s_1 \ge s_1 \ge s_2 \ge s_2 \ge \cdots \ge s_n \ge s_n \ge 0$. *The set* $Re\,\widetilde{D}_n(A) = \{Re\,d : d \in \widetilde{D}_n(A)\}$ *is a convex set in* $\mathbb{R}^n$ *and* $x \in Re\,\widetilde{D}_n(A)$ *if and only if* $|x| \prec_w s$. *Similar characterization is also valid for the real part of the set* $\widetilde{D}_p(A)$, $1 \le p < n$.

COROLLARY 2.12. *The projections of the sets in Corollaries* 2.9 *and* 2.10 *onto* $\mathbb{R}^n$ *are completely characterized by* $|x| \prec_w s$.

3. Inequalities for singular values of certain submatrices. In this section, we establish some inequalities for the singular values of a real or complex skew-symmetric matrices and its right-top submatrix. The following is an application of Theorem 1 of [10].

THEOREM 3.1. 1. *Let*

$$A = \begin{pmatrix} F & B \\ -B^T & G \end{pmatrix}$$

be a $2n \times 2n$ *real skew-symmetric matrix, where* $F, G \in \mathbb{R}_{n\times n}$. *Let* $\alpha_1 \geq \alpha_1 \geq \alpha_2 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq \alpha_n$ *and* $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_n$ *be the singular values of* A *and* B, *respectively.*

(a) *Suppose that* $O^T AO = \begin{pmatrix} 0 & A_1 \\ -A_1 & 0 \end{pmatrix}$ *where* $A_1 = diag\,(\alpha_1, \ldots, \alpha_n)$, *for some* $O \in SO(2n)$, *or equivalently, the canonical form of* A *is*

$$\begin{cases} \begin{pmatrix} 0 & \alpha_1 \\ -\alpha_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & \alpha_n \\ -\alpha_n & 0 \end{pmatrix} & \text{if } [n/2] \text{ is even,} \\ \begin{pmatrix} 0 & \alpha_1 \\ -\alpha_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & \alpha_{n-1} \\ -\alpha_{n-1} & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & -\alpha_n \\ -\alpha_n & 0 \end{pmatrix} & \text{if } [n/2] \text{ is odd.} \end{cases}$$

Then

$$\sum_{i=1}^{k} \beta_i \leq \sum_{i=1}^{k} \alpha_i, \qquad k = 1, \ldots, n,$$
$$\sum_{i=1}^{n-1} \beta_i - \beta_n \leq \sum_{i=1}^{n-1} \alpha_i - \alpha_n,$$

and, in addition, if $\det B < 0$,

$$\sum_{i=1}^{n} \beta_i \leq \sum_{i=1}^{n-1} \alpha_i - \alpha_n.$$

(b) *Suppose that* $O^T AO = \begin{pmatrix} 0 & A_2 \\ -A_2 & 0 \end{pmatrix}$ *where* $A_2 = diag\,(\alpha_1, \ldots, \alpha_{n-1}, -\alpha_n)$, *for some* $O \in SO(2n)$, *or equivalently, the canonical form of* A *is*

$$\begin{cases} \begin{pmatrix} 0 & \alpha_1 \\ -\alpha_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & \alpha_n \\ -\alpha_n & 0 \end{pmatrix} & \text{if } [n/2] \text{ is odd} \\ \begin{pmatrix} 0 & \alpha_1 \\ -\alpha_1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & \alpha_{n-1} \\ -\alpha_{n-1} & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & -\alpha_n \\ \alpha_n & 0 \end{pmatrix} & \text{if } [n/2] \text{ is even.} \end{cases}$$

Then

$$\sum_{i=1}^{k} \beta_i \leq \sum_{i=1}^{k} \alpha_i, \qquad k = 1, \ldots, n,$$
$$\sum_{i=1}^{n-1} \beta_i - \beta_n \leq \sum_{i=1}^{n-1} \alpha_i - \alpha_n,$$

and, in addition, if $\det B > 0$,

$$\sum_{i=1}^{n} \beta_i \le \sum_{i=1}^{n-1} \alpha_i - \alpha_n.$$

2. *Let*

$$A = \begin{pmatrix} F & B \\ -B^T & G \end{pmatrix}$$

be an $(2n+1) \times (2n+1)$ *real skew-symmetric matrix, where* $F \in \mathbb{R}_{n\times n}$ *and* $G \in \mathbb{R}_{(n+1)\times(n+1)}$. *Let* $\alpha_1 \ge \alpha_1 \ge \alpha_2 \ge \alpha_2 \ge \cdots \ge \alpha_n \ge \alpha_n \ge 0$ *and* $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_n \ge 0$ *be the singular values of* A *and* B, *respectively. Then* $\sum_{i=1}^{k} \beta_i \le \sum_{i=1}^{k} \alpha_i$, $k = 1, \ldots, n$.

Proof.

1. Suppose that $m = 2n$. (a) By Corollary 2 in [10], the set

$$\{\operatorname{diag} \begin{pmatrix} U & W \\ V & X \end{pmatrix}^T \begin{pmatrix} 0 & A_1 \\ -A_1 & 0 \end{pmatrix} \begin{pmatrix} U & W \\ V & X \end{pmatrix} [1, \ldots, n | n+1, \ldots, 2n] : \begin{pmatrix} U & W \\ V & X \end{pmatrix} \in SO(2n)\}$$

is the convex hull of the elements $SG\hat{(}n)\alpha$ (see the notation in [10]) where $\alpha = (\alpha_1, \ldots, \alpha_n)$. Now under the hypothesis, A can be assumed to be $\begin{pmatrix} 0 & A_1 \\ -A_1 & 0 \end{pmatrix}$. Moreover, there exist $U, V \in SO(n)$ such that $U^T B V = \operatorname{diag}(\beta_1, \ldots, \beta_n)$ is the singular value decomposition of B if $\det B > 0$. But if $\det B < 0$, we have $U^T B V = \operatorname{diag}(\beta_1, \ldots, \beta_{n-1}, -\beta_n)$. Consider

$$\begin{pmatrix} U^T F U & U^T B V \\ -V^T B^T U & V^T G V \end{pmatrix} = \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}^T \begin{pmatrix} F & B \\ -B^T & G \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}.$$

Now $(\beta_1, \ldots, \beta_n) \in SG\hat{(}n)\alpha$ if $\det B > 0$ and $(\beta_1, \ldots, \beta_{n-1}, -\beta_n) \in SG\hat{(}n)\alpha$ if $\det B < 0$. Then the result follows by using Theorem 4 in [10]. (b) Similar to the first case.

2. Use Corollary 1 and Theorem 4 in [10]. □

THEOREM 3.2. 1. *Let*

$$A = \begin{pmatrix} F & B \\ -B^T & G \end{pmatrix}$$

be a $2n \times 2n$ *complex skew-symmetric matrix, where* $F, G \in \mathbb{C}_{n\times n}$. *Let* $\alpha_1 \ge \alpha_1 \ge \alpha_2 \ge \alpha_2 \ge \cdots \ge \alpha_n \ge \alpha_n$ *and* $\beta_1 \ge \beta_2 \ge \cdots \ge \beta_n$ *be the singular values of* A *and* B, *respectively. Then*

$$\sum_{i=1}^{k} \beta_i \le \sum_{i=1}^{k} \alpha_i, \qquad k = 1, \ldots, n,$$

$$\sum_{i=1}^{n-1} \beta_i - \beta_n \le \sum_{i=1}^{n-1} \alpha_i - \alpha_n.$$

2. *Let*

$$A = \begin{pmatrix} F & B \\ -B^T & G \end{pmatrix}$$

be an $(2n+1)\times(2n+1)$ *complex skew-symmetric matrix, where* $F\in\mathbb{C}_{n\times n}$ *and* $G\in\mathbb{C}_{(n+1)\times(n+1)}$. *Let* $\alpha_1\ge\alpha_1\ge\alpha_2\ge\alpha_2\ge\cdots\ge\alpha_n\ge\alpha_n\ge 0$ *and* $\beta_1\ge\beta_2\ge\cdots\ge\beta_n\ge 0$ *be the singular values of* A *and* B, *respectively. Then* $\sum_{i=1}^k\beta_i\le\sum_{i=1}^k\alpha_i$, $k=1,\dots,n$.

Proof. We only prove the first part of Theorem 3.2 and the proof of the second part is similar. Suppose that the singular values of

$$A=\begin{pmatrix} F & B\\ -B^T & G\end{pmatrix}$$

are $s_1,s_1,s_2,s_2,\dots,s_n,s_n$ and the singular values of B are $\beta_1,\beta_2,\dots,\beta_n$. Let $U^TBV=\operatorname{diag}(\beta_1,\dots,\beta_n)$ be the singular value decomposition of B. Then consider

$$\begin{pmatrix} U^TFU & U^TBV\\ -V^TB^TU & V^TGV\end{pmatrix}=\begin{pmatrix} U & 0\\ 0 & V\end{pmatrix}^T\begin{pmatrix} F & B\\ -B^T & G\end{pmatrix}\begin{pmatrix} U & 0\\ 0 & V\end{pmatrix}.$$

So by Theorem 2.1, $(\beta_1,\beta_2,\dots,\beta_n)\prec_w(s_1,s_2,\dots,s_n)$ and $\sum_{i=1}^{n-1}\beta_i-\beta_n\le\sum_{i=1}^{n-1}s_i-s_n$. ∎

Remark. Regarding Theorem 3.1 and 3.2, the inequalities $\sum_{i=1}^k\beta_i\le\sum_{i=1}^k\alpha_i$, $k=1,\dots,n$ are well known. See [17] for related results. However, the inequalities $\sum_{i=1}^{n-1}\beta_i-\beta_n\le\sum_{i=1}^{n-1}\alpha_i-\alpha_n$ and $\sum_{i=1}^{n}\beta_i\le\sum_{i=1}^{n-1}\alpha_i-\alpha_n$, are new.

4. Some discussions. The following corollary consists of some consequences of Schur–Horn's result, Theorem 1 of [10], Theorems 2.1 and 2.2, respectively.

COROLLARY 4.1. 1. *Let* A *and* C *be* $n\times n$ *real symmetric matrices (Hermitian matrices) with eigenvalues* $\alpha_1\ge\cdots\ge\alpha_n$ *and* $\gamma_1\ge\cdots\ge\gamma_n$, *respectively. Then the set* $\{trAU^*CU : U\in G\}$, *where* $G=SO(n)$ $(U(n))$, *is the interval* $[\sum_{i=1}^n\alpha_i\gamma_{n-i+1},\sum_{i=1}^n\alpha_i\gamma_i]$.

2. [12] *Let* A *and* C *be* $m\times m$ *real skew-symmetric matrices with singular values* $\alpha_1\ge\alpha_1\ge\alpha_2\ge\alpha_2\ge\cdots\ge\alpha_n\ge\alpha_n$ *and* $\gamma_1\ge\gamma_1\ge\gamma_2\ge\gamma_2\ge\cdots\ge\gamma_n\ge\gamma_n$, *respectively.* (i) *When* $m=2n$, *let the canonical forms of* A *and* C *be*

$$\begin{pmatrix} 0 & \alpha_1\\ -\alpha_1 & 0\end{pmatrix}\oplus\cdots\oplus\begin{pmatrix} 0 & \alpha_{n-1}\\ -\alpha_{n-1} & 0\end{pmatrix}\oplus\begin{pmatrix} 0 & \alpha'_n\\ -\alpha'_n & 0\end{pmatrix}$$

and

$$\begin{pmatrix} 0 & \gamma_1\\ -\gamma_1 & 0\end{pmatrix}\oplus\cdots\oplus\begin{pmatrix} 0 & \gamma_{n-1}\\ -\gamma_{n-1} & 0\end{pmatrix}\oplus\begin{pmatrix} 0 & \gamma'_n\\ -\gamma'_n & 0\end{pmatrix},$$

respectively. Here $\alpha'_n=\pm\alpha_n$ *and* $\gamma'_n=\pm\gamma_n$. *Then the set* $\{trAU^TCU : U\in SO(m)\}$ *is the interval* $[a,b]$ *where* $a=-2(\sum_{i=1}^{n-1}\alpha_i\gamma_i+\alpha'_n\gamma'_n)$ *and*

$$b=\begin{cases} 2(\sum_{i=1}^{n-1}\alpha_i\gamma_i+\alpha'_n\gamma'_n) & \text{if } n \text{ is even,}\\ 2(\sum_{i=1}^{n-1}\alpha_i\gamma_i-\alpha'_n\gamma'_n) & \text{if } n \text{ is odd.}\end{cases}$$

(ii) *When* $m=2n+1$, *the set* $\{trAU^TCU : U\in SO(m)\}$ *is the interval*

$$\left[-2\sum_{i=1}^n\alpha_i\gamma_i,2\sum_{i=1}^n\alpha_i\gamma_i\right].$$

3. [9] *Let A and C be an $m \times m$ complex skew-symmetric matrices with singular values $\alpha_1 \geq \alpha_1 \geq \alpha_2 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq \alpha_n$ and $\gamma_1 \geq \gamma_1 \geq \gamma_2 \geq \gamma_2 \geq \cdots \geq \gamma_n \geq \gamma_n$, respectively, where $n = [m/2]$. Then the congruence numerical range $W_A(C) = \{trAU^TCU : U \in U(n)\}$ is a circular disk of radius $2\sum_{i=1} \alpha_i\gamma_i$ and centered at the origin.*

Proof. We only deal with the last part. The set $W_A(C)$ has circular symmetry and the radius is obtained by Theorem 2.1 and 2.2. The set $W_A(C)$ can be reformulated as

$$W_A(C) = \begin{cases} \left\{ \text{tr} \begin{pmatrix} 0 & A \\ -A & 0 \end{pmatrix} U^T \begin{pmatrix} 0 & \Gamma \\ -\Gamma & 0 \end{pmatrix} U : U \in U(2n) \right\} & if \text{ m = 2n,} \\ \left\{ \text{tr} \left[\begin{pmatrix} 0 & A \\ -A & 0 \end{pmatrix} \oplus 0 \right] U^T \left[\begin{pmatrix} 0 & \Gamma \\ -\Gamma & 0 \end{pmatrix} \oplus 0 \right] U : U \in U(2n+1) \right\} & \\ & if \text{ m = 2n+1.} \end{cases}$$

Here $A = \text{diag}\,(\alpha_1, \ldots, \alpha_n)$ and $\Gamma = \text{diag}\,(\gamma_1, \ldots, \gamma_n)$. In other words, $W_A(C) = \{-2\alpha \cdot d : d \in \widetilde{D}_n\,(\Gamma_{n,n})\}$ where $n = [m/2]$, $\alpha = (\alpha_1, \ldots, \alpha_n)$, and

$$\Gamma_{n,n} = \begin{pmatrix} 0 & \Gamma \\ -\Gamma & 0 \end{pmatrix}.$$

The origin is contained in $W_A(C)$. It is because the zero vector is an element of $\widetilde{D}_n\,(\Gamma_{n,n})$ by Theorem 2.1 and 2.2. □

Remark. 1. The upper bound $\sum_{i=1}^n \alpha_i\gamma_i$ in the first part was obtained by Fan [2]. Also see [13, 4].

2. The intervals given in the first and second part of Corollary 4.1 are useful for plotting generalized numerical ranges [12, 5]

3. Regarding the last part of the above result, Choi et al. [1] used the simply-connectedness of $SU(n)$ with a homotopic argument to prove that the origin is contained in $W_A(C)$ even if A and C are complex matrices.

Now if $K \subset \mathbb{R}^p$, we define $\mathcal{T}(K)$ to be the torus generated by K, i.e.,

$$\mathcal{T}(K) = \{(e^{i\theta_1}d_1, \ldots, e^{i\theta_n}d_n) : (d_1, \ldots, d_n) \in K,\ \theta_1, \ldots, \theta_n \in \mathbb{R}\}.$$

Let A be an $m \times m$ complex skew-symmetric matrix. Then $\widetilde{D}_p(A) = \widetilde{D}_p(S_{n,n}) = \mathcal{T}(D_p(S_{n,n}))$ in view of Theorem 1 of [10], Theorems 2.1 and 2.2. In other words, if we denote by $D(n) \subset U(n)$ the subgroup of diagonal matrices, then

$$\begin{aligned} \widetilde{D}_p(S_{n,n}) &\equiv \{\text{diag}\,(U^TSU)[1, \ldots, p|n+1, \ldots, n+p] : U \in U(m)\} \\ &= \{\text{diag}\,((UD)^TS(UD))[1, \ldots, p|n+1, \ldots, n+p] : U \in SO(m), \\ &\qquad D = D_1 \oplus D_1(\oplus 1), D_1 \in D(n)\}. \end{aligned}$$

Regarding Thompson–Sing's result, we have the same phenomenon [14, 11] (see the description of the set $\{\text{diag}\,(USV) : U, V \in SO(n)\}$ [14] and Theorem 1 of [10]).

So we conclude that

$$\begin{aligned} \{\text{diag}\,(USV) : U, V \in U(n)\} &= \mathcal{T}(\{\text{diag}\,(USV) : U, V \in SO(n)\}) \\ &= \{\text{diag}\,(DUSV) : U, V \in SO(n), D \in D(n)\} \\ &= \{\text{diag}\,(DUSV) : U, V \in SO(n), D \in D(n)\} \\ &= \{\text{diag}\,(D_1USD_2V) : U, V \in SO(n), D_1, D_2 \in D(n)\}. \end{aligned}$$

Hence, the subset $[D(n) \oplus D(n)(\oplus 1)]SO(m) \subset U(m)$ generates $\widetilde{D}_n$ $(S_{n,n})$, $n = [m/2]$ and $D(n)SO(n) \subset U(n)$ generates $\{\mathrm{diag}\,(USV) : U, V \in U(n)\}$.

On the other hand, no torus relationship exists for the real and complex symmetric cases, i.e., $\{\mathrm{diag}\,(U^T SU) : U \in U(n)\} \neq \mathcal{T}(\{\mathrm{diag}\,(U^T SU) : U \in SO(n)\})$ in view of Theorem 1 of [16] and Schur–Horn's result. See the corollary in [16] for unitary symmetric matrices.

The following results can be viewed as corollaries of Schur–Horn's result. The first part is for the diagonal elements and singular values of Hermitian (real symmetric) matrices. The second part is particularly for the symmetric special orthogonal matrices. We denote by $\hat{B}$ the convex hull of the set B and Σ_n, the symmetric group on $\{1, \ldots, n\}$.

COROLLARY 4.2. 1. *There exists a Hermitian (real symmetric) matrix with prescribed diagonal element $d = (d_1, \ldots, d_n) \in \mathbb{R}^n$ and singular values $s_1 \geq \cdots \geq s_n$ if and only if $d \prec (\pm s_1, \ldots, \pm s_n)$ for some choice of signs. So the set of diagonal elements of a Hermitan (real symmetric) matrix with prescribed singular values s's is $\cup_{0 \leq k \leq n} \overset{\wedge}{\Sigma_n s}(k)$ where $s(k) = (-s_1, \ldots, -s_k, s_{k+1}, \ldots, s_n)$.*

2. *There exists a symmetric special orthogonal matrix with prescribed diagonal element $d = (d_1, \ldots, d_n)$ if and only if $d \prec (\pm 1, \ldots, \pm 1)$ for some choice of signs such that the number of negative terms is even. So the set of diagonal elements of a symmetric special orthogonal matrix is $\cup_{0 \leq i \leq [n/2]} \overset{\wedge}{\Sigma_n e}(2k)$ where $e(2k)$ is a vector whose first $2k$ entries are -1 and the remaining entries are 1, $0 \leq k \leq [n/2]$.*

Proof. (1) A Hermitan (real symmetric) matrix A has the spectral decomposition $U^* AU = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)$ where λ's are real. Evidently, λ's are $\pm s$'s so by Schur–Horn's result, $d \prec \lambda = (\pm s_1, \ldots, \pm s_n)$, for some choice of signs and vice versa. (2) The singular values of a symmetric special orthogonal matrix are 1's and the eigenvalues are ± 1 where number of -1 terms is even. □

Remark. The set $\cup_{0 \leq k \leq n} \overset{\wedge}{\Sigma_n s}(k)$ where $s(k) = (-s_1, \ldots, -s_k, s_{k+1}, \ldots, s_n)$ is different from $\mathcal{W}_{\mathbb{R}} s$. The former one is not convex in general but the later one is always convex. When $n = 2$, $\cup_{0 \leq k \leq n} \overset{\wedge}{\Sigma_n s}(k)$ is the union of the line segments determined by the order pairs (s_1, s_2), (s_2, s_1); $(-s_1, s_2)$, $(s_2, -s_1)$; $(s_1, -s_2)$, $(-s_2, s_1)$; $(-s_1, -s_2)$, $(-s_2, -s_1)$. But the set $\mathcal{W}_{\mathbb{R}} s$ is the convex hull of all the points. When $n = 3$, $\cup_{0 \leq k \leq n} \overset{\wedge}{\Sigma_n s}(k)$ is the union of eight hexagons and $\mathcal{W}_{\mathbb{R}} s$ is the convex hull of all those hexagons. In general, $\cup_{0 \leq k \leq n} \overset{\wedge}{\Sigma_n s}(k) \subset \mathcal{W}_{\mathbb{R}} s$ and the convex hull of $\cup_{0 \leq k \leq n} \Sigma_n s(k)$ is $\mathcal{W}_{\mathbb{R}} s$ since $\mathcal{W}_{\mathbb{R}} s$ is $\overset{\wedge}{G(n)} s$.

Finally, we state the result for skew-symmetric unitary matrices and skew-symmetric special orthogonal matrices. I hope that physicists can find applications of the results as they did for symmetric unitary matrices regarding S-matrix (symmetric unitary), which describes a system of two-body reactions $A_i + B_i \to A_j + B_j$, $i, j = 1, \ldots, n$ for fixed energy and angular momentum [18].

COROLLARY 4.3. *The partial diagonal elements $a = (a_{12}, a_{34}, \ldots, a_{2n-1,2n})$ of an $m \times m$ ($n = [m/2]$) skew-symmetric unitary matrix A are completely described by, after rearranging the entries of a in descending order with respect to modulus,*

1. *when $m = 2n$, $\max_{1 \leq i \leq n} |a_{2i-1,2i}| \leq 1$, and $\sum_{i=1}^{n-1} |a_{2i-1,2i}| - |a_{2n-1,n}| \leq n - 2$.*
2. *when $m = 2n + 1$, $\max_{1 \leq i \leq n} |a_{2i-1,2i}| \leq 1$.*

Proof. The singular values of a unitary matrix are 1's. Then apply Theorems 2.1

and 2.2. The inequalities $\sum_{i=1}^{k} |a_{2i-1,2i}| \le k$, $k = 1, \dots, n$, amount to $|a_{12}| \le 1$, after rearranging the entries of a in descending order with respect to modulus. □

COROLLARY 4.4. *The partial diagonal elements $a = (a_{12}, a_{34}, \dots, a_{2n-1,2n})$ of an $m \times m$ ($n = [m/2]$) skew-symmetric special orthogonal matrix A are completely described by, after rearranging the entries of a in descending order with respect to absolute value,*

1. *when $m = 2n$, and*

(a) *if the canonical form of A is* $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$,

$$\max_{1 \le i \le n} |a_{2i-1,2i}| \le 1,$$

$$\sum_{i=1}^{n-1} |a_{2i-1,2i}| - |a_{2n-1,2n}| \le n - 2,$$

and, in addition, if the number of negative terms among a is odd,

$$\sum_{i=1}^{n} |a_{2i-1,2i}| \le n - 2.$$

(b) *if the canonical form of A is* $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \cdots \oplus \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \oplus \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$,

$$\max_{1 \le i \le n} |a_{2i-1,2i}| \le 1,$$

$$\sum_{i=1}^{n-1} |a_{2i-1,2i}| - |a_{2n-1,2n}| \le n - 2,$$

and, in addition, if the number of negative terms among a is even,

$$\sum_{i=1}^{n} |a_{2i-1,2i}| \le n - 2.$$

2. *when $m = 2n + 1$,* $\max_{1 \le i \le n} |a_{2i-1,2i}| \le 1$.

Proof. The singular values of an orthogonal matrix are 1's. Then apply Theorem 1 of [10]. The inequalities $\sum_{i=1}^{k} |a_{2i-1,2i}| \le k$, $k = 1, \dots, n$, amount to $|a_{12}| \le 1$, after rearranging the entries of a in descending order. □

COROLLARY 4.5. *The partial diagonal elements $a = (a_{12}, a_{34}, \dots, a_{2n-1,2n})$ of an $m \times m$ ($n = [m/2]$) skew-symmetric orthogonal matrix A are completely described by, after rearranging the entries of a in descending order with respect to absolute value,*

1. *when $m = 2n$,*

$$\max_{1 \le i \le n} |a_{2i-1,2i}| \le 1,$$

$$\sum_{i=1}^{n-1} |a_{2i-1,2i}| - |a_{2n-1,2n}| \le n - 2.$$

2. *when $m = 2n + 1$,* $\max_{1 \le i \le n} |a_{2i-1,2i}| \le 1$.

5. A list. In [15] Thompson listed some current results, in terms of Hadamard product, on the diagonal elements, eigenvalues, and singular values of various matrices. We want to enlarge the list so that readers can browse the results. We will use the notation $|d| \dashv s$ to denote the relationship given by 1 and 2, $|d| \lhd s$ for (1), (2), and (3) in [16] and $d \ll s$ for (15), (16), and (17) in [10]. Whenever the real part is seen, imaginary parts work, too. One can consider the projection of the underlying set onto $e^{i\theta}\mathbb{R}^n$ where $\theta \in \mathbb{R}$ is fixed and the corresponding result is also valid.

1. When is $x = Sy$ $(x, y \in \mathbb{R}^n)$ with S doubly stochastic? ($x \prec y$, Hardy–Littlewood–Polya).
2. When is $x = Sy$ $(x, y \in \mathbb{R}^n)$ S unistochastic ($S = U \circ \overline{U}$, U unitary) or orthostochastic ($S = O \circ O$, O orthogonal)? ($x \prec y$, Schur–Horn).[1]
3. When is $x = Sy$ with $S \in \mathcal{W}_{\mathbb{C}}$, ($S \in \mathcal{W}_{\mathbb{R}}$ and $x, y \in \mathbb{R}^n$) absolutely substochastic? ($|x| \prec_w |y|$, Thompson, 1977)
4. When is $x = Sy$ with $S = U \circ V$ where U, V are unitary? ($|x| \dashv |y|$, Thompson–Sing, 1977)
5. When is $x = \operatorname{Re} Sy$ $(x, y \in \mathbb{R}^n)$ with $S = U \circ V$ where U, V are unitary? ($|x| \prec_w |y|$, Tam)
6. When is $x = Sy$ $(x, y \in \mathbb{R}^n)$ with $S = U \circ V$ where U, V are special orthogonal? ($x \ll y$, i.e., we have inequality (17) in [10] if the total number of negative terms among x and y is odd, Thompson 1977 when $y \in \mathbb{R}^n_+$, [10] for arbitrary $y \in \mathbb{R}^n$)
7. When is $x = Sy$ $(x, y \in \mathbb{R}^n)$ ith $S = U \circ V$ where U, V are orthogonal? ($|x| \dashv |y|$, Thompson 1977 when $y \in \mathbb{R}^n_+$, [10] for arbitrary $y \in \mathbb{R}^n$)
8. When is $x = Sy$ with $S = U \circ U$ where U is unitary? ($|x| \lhd |y|$, Thompson 1979)
9. When is $x = \operatorname{Re} Sy$ with $S = U \circ U$ where U is unitary? ($|x| \prec_w |y|$, [11])
10. When is $x = Sy$ with $S = U \circ X - V \circ W$ where $\begin{pmatrix} U & V \\ W & X \end{pmatrix}$ is unitary? ($|x| \dashv |y|$)
11. When is $x = \operatorname{Re} Sy$ $(x, y \in \mathbb{R}^n)$ with $S = U \circ X - V \circ W$ where $\begin{pmatrix} U & V \\ W & X \end{pmatrix}$ is unitary? ($|x| \prec_w |y|$)
12. When is $x = Sy$ $(x, y \in \mathbb{R}^n)$ with $S = U \circ X - V \circ W$ where $\begin{pmatrix} U & V \\ W & X \end{pmatrix}$ is special orthogonal? ($x \ll y$ [10])
13. When is $x = Sy$ $(x, y \in \mathbb{R}^n)$ with $S = U \circ X - V \circ W$ where $\begin{pmatrix} U & V \\ W & X \end{pmatrix}$ is orthogonal? ($|x| \dashv |y|$ [10])
14. When is $x = Sy$ or $x = \operatorname{Re} Sy$ $(x, y \in \mathbb{R}^n)$ with $S = U \circ X - V \circ W$ where Z is unitary (special unitary, and $x, y \in \mathbb{R}^n$ while Z is orthogonal or special orthogonal) with $Z[1, \ldots, 2n | 1, \ldots, 2n] = \begin{pmatrix} U & V \\ W & X \end{pmatrix}$? ($|x| \prec_w |y|$)

REFERENCES

[1] M. D. Choi, C. Laurie, H. Radjavi, and P. Rosenthal, *On the congruence numerical range and related functions of matrices*, Linear and Multilinear Algebra, 22 (187), pp. 1–5.

[1] The result is true when $U(n)$ is replaced by $SU(n)$ and $O(n)$ is replaced by $SO(n)$. The corresponding matrices can be called special unistochastic and special orthostochastic, if you like.

[2] K. Fan, *On a theorem of Weyl concerning eigenvalues of linear transformations*, Proc. Nat. Acad. Sci. U.S.A., 35 (1949), pp. 652–655.
[3] A. Horn, *Doubly stochastic matrices and the diagonal of a rotation matric*, Amer. J. Math., 76 (1954), pp. 620–630.
[4] A. S. Lewis, *Von Neumann's Lemma and a Chevalley-Type Theorem for Convex Functions on Cartan Subspaces*, Combinatorics and Optimization Research Report, University of Waterloo, Canada, 1995.
[5] C. K. Li, C. H. Sung, and N. K. Tsing, *The c-convex matrices: Characterizations, inclusion relations and normality*, Linear and Multilinear Algebra, 25 (1989), pp. 275–287.
[6] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.
[7] A. L. Onishchik and E. B. Vinberg, *Lie Groups and Algebraic Groups*, Springer-Verlag, Berlin, 1990.
[8] F. Y. Sing, *Some results on matrices with prescribed diagonal elements and singular values*, Canad. Math. Bull., 19 (1976), pp. 89–92.
[9] T. Y. Tam, *Note on a paper of R.C. Thompson: "The congruence numerical range"*, Linear and Multilinear Algebra, 17 (1985), pp. 107–115.
[10] T. Y. Tam, *Kostant's convexity theorem and the compact classical groups*, Linear and Multilinear Algebra, to appear.
[11] T. Y. Tam, *A Lie Theoretic Approach of Thompson's Theorems on Singular Values-Diagonal Elements and Some Related Results*, manuscript, 1996.
[12] T. Y. Tam, *Plotting the Generalized Numerical Ranges Associated with the Compact Classical Groups*, manuscript, 1996.
[13] C. M. Theobald, *An inequality for the trace of the product of two symmetric matrices*, Math. Proc. Cambridge Philos. Soc., 77 (1975), pp. 265–267.
[14] R. C. Thompson, *Singular values, diagonal elements, and convexity*, SIAM J. Appl. Math., 32 (1977), pp. 39–63.
[15] R. C. Thompson, *Diagonal elements, singular values, majorization, reflection groups, elementary divisors*, Johns Hopkins Lecture, 4.
[16] R. C. Thompson, *Singular values and diagonal elements of complex symmetric matrices*, Linear Algebra Appl., 26 (1979), pp. 65–106.
[17] R. C. Thompson, *Singular value inequalities for matrix sums and minors*, Linear Algebra Appl., 11 (1975), pp. 251–269.
[18] S. Waldenstrom, *S-matrix and unitary bounds for three channel systems, with applications to low-energy photoproduction of pions from nucleons*, Nuclear Phys. B, 77 (1974), pp. 479–493.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 755–771, July 1998

010

EXTENDED KRYLOV SUBSPACES: APPROXIMATION OF THE MATRIX SQUARE ROOT AND RELATED FUNCTIONS*

VLADIMIR DRUSKIN† AND LEONID KNIZHNERMAN‡

Abstract. We introduce an economical Gram–Schmidt orthogonalization on the extended Krylov subspace originated by actions of a symmetric matrix and its inverse. An error bound for a family of problems arising from the elliptic method of lines is derived. The bound shows that, for the same approximation quality, the diagonal variant of the extended subspaces requires about the square root of the dimension of the standard Krylov subspaces using only positive or negative matrix powers. An example of an application to the solution of a 2.5-D elliptic problem attests to the computational efficiency of the method for large-scale problems.

Key words. Krylov subspace, Lanczos method, rational approximations, functions of matrices, matrix square root, method of lines, elliptic problems, 2.5-D problems

AMS subject classifications. 15A15, 65N40, 65F10

PII. S0895479895292400

1. Introduction.

Let us consider computation of the vector

$$u = f(A)\varphi, \tag{1.1}$$

where A is a real symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, φ is a nonzero vector from $\mathbf{R}^n$, and f is a function analytic on $[\lambda_1, \lambda_n]$. There exists a method to approximate u by an element of a Krylov subspace

$$\mathcal{K}^m(A, \varphi) = \text{ span } \left\{\varphi, A\varphi, \ldots, A^{m-1}\varphi \quad ,\right.$$

which is called the Spectral Lanczos Decomposition Method (SLDM). This method involves execution of the m steps of the Lanczos algorithm [20, ch. 13] with A and φ, providing the Lanczos vectors $q_1, \ldots, q_m$ and tridiagonal symmetric $m \times m$ matrix H, and takes as an approximant to u the vector

$$u_m = \|\varphi\|(q_1, \ldots, q_m) f(H) e_1^{(m)},$$

with $e_1^{(m)}$ the first unit m-vector. This technique appeared in the literature from the mid-eighties [15, 17, 19, 28, 3, 8, 10], and it can now be viewed as standard. It is sometimes possible to calculate Krylov subspaces, originated by A^{-1}. The simplest example is the case of a 1-D elliptic operator A (see [17]). Further, if A is a finite difference approximation of a 2-D elliptic operator (e.g., from the 2.5-D direct current problem), LU decomposition of A requires $O\ \ n^{1.5}\big)$ arithmetical operations (that should be done only once), while the computation of each action of A^{-1} on a vector takes only $O(n \log n)$ operations by the nested dissection method [7]; it is not much larger than $O(n)$ operations required for the computation of action of A. For some important problems (Maxwell's system, acoustic equations) A can be expressed as

*Received by the editors September 25, 1995; accepted for publication (in revised form) by B. Kågström May 15, 1997; published electronically April 2, 1998.
http://www.siam.org/journals/simax/19-3/29240.html
†Schlumberger-Doll Research, Old Quarry Road, Ridgefield, CT 06877-4108 (druskin@ridgefield.sdr.slb.com).
‡Central Geophysical Expedition, Narodnogo Opolcheniya St., House 40, Bldg. 3, Moscow 123298, Russia (knizhner@cgemath.msk.su).

the product of diagonal and Laplace operators, so A^{-1} can be applied by fast Fourier transform (FFT) for $O(n \log n)$ operations even without explicit inversion.

This observation has led us to use of "extended" Krylov subspaces

$$\mathcal{K}^{k,m} = \mathcal{K}^{k,m}(A,\varphi) = \text{ span } \left\{A^{-k+1}\varphi, \ldots, A^{-1}\varphi, \varphi, \ldots, A^{m-1}\varphi \right. \quad ,$$
$$m \geq 1, \quad k \geq 1, \qquad \dim \mathcal{K}^{k,m} \leq k+m-1.$$

Evidently, $\mathcal{K}^{k,m}(A,\varphi) = \mathcal{K}^{k+m-1}(A, A^{-k+1}\varphi)$ can be treated as a usual Krylov subspace, but with another starting vector — $A^{-k+1}\varphi$. The term "extended Krylov subspace" re ects the fact that we do not wish to calculate $A^{-k+1}\varphi$ (it would be numerically unstable), but instead are interested in developing a special procedure to obtain an orthonormal basis of $\mathcal{K}^{k,m}$ and to solve the corresponding Ritz problem. This procedure can be interpreted as "extending" $\mathcal{K}^m$ with negative powers of A and will be called the Extended Krylov Subspace Method (EKSM).

The Extended Krylov Subspaces can be considered a special case of the Rational Krylov Subspaces with multiple shifts, suggested by [23, 24, 25] for the nonsymmetric eigenproblems. However, $\mathcal{K}^{k,m}$ will allow us to construct a *short orthogonalization recursion* with up to four terms, compared with the *full Gram–Schmidt orthogonalization* considered by Ruhe. An interesting scheme suggested by [9] for application to control theory reduces the number of computed inner products compared to [25], but it uses bi-orthogonalization even for the symmetric problems, i.e., it looses the optimality of the Galerkin method and does not allow one to make a priori error estimates.

We consider problems arising from a solution of elliptic equations by the method of lines. This method results in boundary value problems for the equation

$$Aw - \frac{d^2 w}{d\theta^2} = g(\theta)\varphi. \tag{1.2}$$

The solution of (1.2) can be presented as $w(\theta) = f(A)\varphi$, with $f(x) = \sqrt{x}$, or $f(x) = e^{-\theta\sqrt{x}}$, or a rational function of these two functions. Starting from [27], SLDM was used for solving (1.2) by many applied researchers [13, 22, 14, 1, 29]. The error bound given in [4] shows that SLDM applied to (1.2) has the same logarithmical rate of convergence as conjugate gradients (CG) for linear systems with the matrix A.

In section 2 we introduce functional classes to handle and establish some sufficient conditions for a function to belong to the classes. We also present a few examples.

In section 3 we investigate approximating properties of $\mathcal{K}^{m,m}$.

In section 4 we prove that for functions f from the family being considered the Extended Krylov Subspace Method using $\mathcal{K}^{m,m}$ takes the number of steps approximately equal to the square root of SLDM's one to converge.

We derive in section 5 an economical Gram–Schmidt orthogonalization on $\mathcal{K}^{k,m}$. Its arithmetical and storage requirements (of course, not including the additional cost of matrix factorization and action of A^{-1}) do not exceed the ones of $m+k$ steps of the standard CG [12, section 10.2].

In section 6 we describe the EKSM matrix-functional formula and ways of its realization.

Finally, we present in section 7 an example of an application to a 2.5-dimensional elliptical problem arising in the context of geophysical electrical tomography. A reduction of one order in computational time and two orders in number of iterations compared to the SLDM is observed. Our experiments also indicate that the EKSM clearly benefits from the short orthogonalization recurrence, which is the main new feature compared to the Rational Krylov Subspace algorithm.

2. Classes of functions and examples. In this paper we shall apply EKSM to functions of two classes, one of which contains the other.

We say that a function f, defined on $]0,+\infty[$, belongs to *class* $\mathcal{A}$, if it is presentable in the form

$$f(x) = \int_{-\infty}^{0} (x-\lambda)^{-1}\, d\gamma(\lambda), \qquad x > 0, \tag{2.1}$$

where γ is a positive measure on $]-\infty, 0]$ such that $-\lambda^{-1}\cdot\gamma$ is a bounded measure on $]-\infty,-1]$,[1] i.e.,

$$\int_{-\infty}^{-1} -\lambda^{-1}\big)\, d\gamma(\lambda) < +\infty. \tag{2.2}$$

We define *class* $\mathcal{B}$ as the closure of class $\mathcal{A}$ with respect to taking linear combinations and multiplying by monomials x^l with $l \in \mathbf{Z}$.

In accordance with (2.1), we shall assume that $\lambda_1 > 0$, i.e., the matrix A is positive definite.

A trivial example of a function from class $\mathcal{A}$ is x^{-1}, taking form (2.1) with γ the unit measure concentrated at 0 (δ-function). As a consequence, all the monomials x^l ($l \in \mathbf{Z}$) belong to class $\mathcal{B}$.

We shall see further (in the proof of Theorem 1) that any function from class $\mathcal{A}$ (and, therefore, class $\mathcal{B}$) is analytic in the set $\mathbf{C}\backslash]-\infty, 0]$. Now we shall establish two sufficient conditions for belonging to class $\mathcal{B}$.

PROPOSITION 1. *Let a function f, real-valued for positive real arguments, be analytic in $\mathbf{C}\backslash]-\infty,0]$ and continuous at the upper and lower edges of the cut $]-\infty,0]$ with a possible exception of the point 0. Let f also satisfy the conditions*

$$\sup_{|\zeta|=R,\ \zeta\neq -R} |f(\zeta)| = o\ \ R^{-1}\big) \quad \text{as } R \to +0, \tag{2.3}$$

$$\sup_{|\zeta|=R,\ \zeta\neq -R} |f(\zeta)| = o(1) \quad \text{as } R \to +\infty, \tag{2.4}$$

$$\int_{-1}^{0} |\Im f(\lambda + 0i)| d\lambda < +\infty, \tag{2.5}$$

$$\int_{-\infty}^{-1} -\lambda^{-1}\big)\, |\Im f(\lambda + 0i)| d\lambda < +\infty. \tag{2.6}$$

Then f belongs to class $\mathcal{B}$.

Proof. Denote by Γ_R the circumference of radius R, centered at 0 and oriented in the positive direction. For $0 < r < |z| < R$ we obtain, owing to Cauchy's formula

$$f(z) = \frac{1}{2\pi i}\int_{\Gamma_R}^{*} \frac{f(\zeta)}{\zeta - z} d\zeta - \frac{1}{2\pi i}\int_{\Gamma_r}^{*} \frac{f(\zeta)}{\zeta - z} d\zeta + \frac{1}{\pi}\int_{-R}^{-r} (\lambda - z)^{-1} \Im f(\lambda + 0i)\, d\lambda \tag{2.7}$$

[1] We could have used any negative number instead of 1 here.

(an asterisk means here that the point $-\rho$ at a contour Γ_ρ is taken twice, allowing the function to attain different values at the two edges of the cut). Conditions (2.3) and (2.4) imply that the integrals on Γ_r and Γ_R in (2.7) vanish as $r \to +0$ and $R \to +\infty$, respectively. So, we get an improper integral representation

$$f(z) = \frac{1}{\pi} \int_{-\infty}^{0} (\lambda - z)^{-1} \Im f(\lambda + 0i)\, d\lambda.$$

Due to (2.5), the function $\pi^{-1}\Im f(\lambda+0i)$ induces a continuous measure γ on $]-\infty, 0]$. Set $\gamma = \gamma^+ - \gamma^-$ with $\gamma^+ = \max(\gamma, 0)$ and $\gamma^- = \max(-\gamma, 0)$. By virtue of (2.6), the restriction of γ^+ and γ^- on $]-\infty, -1]$ are bounded measures. It remains to note that

$$f(z) = \int_{-\infty}^{0} (\lambda - z)^{-1} d\gamma^+(\lambda) - \int_{-\infty}^{0} (\lambda - z)^{-1} d\gamma^-(\lambda)$$

and the measures γ^+ and γ^- are positive. □

PROPOSITION 2. *If*

$$f(z) = \sum_{k=1}^{\infty} r_k (z - z_k)^{-1},$$

with $r_k, z_k \in \mathbf{R}$, $z_k < 0$, $z_k \to -\infty$ *as* $k \to +\infty$ *and*

$$\sum_{k=1}^{\infty} \left| r_k z_k^{-1} \right| < +\infty, \tag{2.8}$$

then the function f *belongs to class* $\mathcal{B}$.

Proof. Evidently, the measure

$$\gamma(\lambda) = \sum_{k=1}^{\infty} r_k \delta(\lambda - z_k)$$

is correctly defined and satisfies (2.1). Condition (2.8) guarantees that γ's positive and negative components γ^+ and γ^- obey (2.2). □

List a few nontrivial examples of functions from class $\mathcal{B}$.

If $w(\theta) = f(A)\varphi$ is the solution of (1.2) for a Fourier-transformable function g and homogeneous boundary conditions, then function f belongs to class $\mathcal{B}$. Here we demonstrate specific examples arising from a solution of elliptic equations by the method of lines.

Example 1. Let us consider

$$w(\theta) = \exp\left(-\theta\sqrt{A}\right)\varphi, \tag{2.9}$$

which is the solution of the boundary value problem $w(0) = \varphi$, $w(+\infty) = 0$ for equation (1.2) with $g = 0$. Then, the corresponding function can be presented as

$$f(x) = e^{-\theta\sqrt{x}} = x\hat{f}(x) + 1, \qquad \hat{f}(x) = \frac{e^{-\theta\sqrt{x}} - 1}{x} = \int_{-\infty}^{0} (x - \lambda)^{-1} \frac{\sin(\theta\sqrt{-\lambda})d\lambda}{\pi\lambda}.$$

Example 2. A more familiar case is the matrix square root (see [2, section 2.2])

$$f(x) = \sqrt{x} = x\hat{f}(x), \qquad \hat{f}(x) = x^{-1/2} = \int_{-\infty}^{0} (x-\lambda)^{-1} \frac{d\lambda}{\pi\sqrt{-\lambda}}.$$

This function arises from the Dirichlet to Neumann mapping problem for (2.9) at $\theta = 0$.

In Examples 1 and 2 the functions $\hat{f}$ and, therefore, f belong to class $\mathcal{B}$ by virtue of Proposition 1. The generating measures are continuous.

Example 3.

$$f(x) = 0.5\left[e^{-\theta\sqrt{x}} + e^{-(2\pi-\theta)\sqrt{x}}\right]\left[\sqrt{x}\left(1 - e^{-2\pi\sqrt{x}}\right)\right]^{-1} = \frac{1}{2\pi}x^{-1} + \frac{1}{\pi}\sum_{k=1}^{\infty} \left. x + k^2\right)^{-1} \cos k\theta. \tag{2.10}$$

This function produces the periodical solution of (1.2) with

$$g = \sum_{k=-\infty}^{\infty} \delta(\theta - 2\pi k)\varphi;$$

it is introduced for the 2.5-D direct current problem in the cylindrical coordinate system. Proposition 2 is applicable to the series in (2.10), and the generating measure is discrete.

3. Approximating properties of diagonal extended Krylov subspaces. To estimate the error of any method based on $\mathcal{K}^{k,m}$, we have first to examine its approximating properties. The following two theorems display superior approximating properties of $\mathcal{K}^{m,m}$ over the standard Krylov subspace $\mathcal{K}^j$ for matrix functions from classes $\mathcal{A}$ and $\mathcal{B}$.

We shall denote by Φ the inverse Zhukovski function $\Phi(z) = z + \sqrt{z^2-1}$.

THEOREM 3. *The error estimate*[2]

$$\min_{p(X)\in\mathbf{R}[X],\ \deg p\le 2m} \left\| f(X) - X^{-m}p(X) \right\|_{C[\lambda_1,\lambda_n]} \lesssim m^2\Phi\left(\frac{\sqrt{\lambda_n}+\sqrt{\lambda_1}}{\sqrt{\lambda_n}-\sqrt{\lambda_1}}\right)^{-m} \tag{3.1}$$

is valid for approximation of any function f from class $\mathcal{A}$.

Proof. Formula (2.1) defines a function $f(z)$ analytic for $z \in \mathbf{C}\backslash]-\infty, 0]$. Really, Markov's function [16, ch. 2, section 6]

$$\int_{-1}^{0} (z-\lambda)^{-1}\, d\gamma(\lambda)$$

is analytic in $\mathbf{C}\backslash]-1, 0]$. On the other hand, for z, belonging to any compactum in $\mathbf{C}\backslash]-\infty, 0]$,

$$\left|\int_a^b (z-\lambda)^{-1}\, d\gamma(\lambda)\right| \le \int_a^b \left|\frac{z}{\lambda} - 1\right| \ -\lambda^{-1}\big)\ d\gamma(\lambda)$$

[2] $\mathbf{R}[X]$ denotes the set of polynomials with real coefficients, deg denotes the degree of a polynomial, and $\|f\|_{C[\lambda_1,\lambda_n]}$ denotes $\max_{[\lambda_1,\lambda_n]}|f|$ for a continuous function f.

$$\lesssim \int_a^b -\lambda^{-1}\big)\, d\gamma(\lambda) \to 0 \quad \text{as } a, b \to -\infty, \quad a < b < -1,$$

due to condition (2.2) (the symbol $\lesssim$ denotes the same as O). This implies that the improper integral

$$\int_{-\infty}^{-1} (z-\lambda)^{-1}\, d\gamma(\lambda)$$

uniformly converges on a compactum and, therefore, is analytic in $\mathbf{C}\backslash]-\infty, 0]$.

Put

$$a = \sqrt{\lambda_1 \lambda_n}, \qquad t = \frac{\sqrt{\lambda_n} + \sqrt{\lambda_1}}{\sqrt{\lambda_n} - \sqrt{\lambda_1}},$$

$$f_1(z) = \int_{-\infty}^{-a} (z-\lambda)^{-1}\, d\gamma(\lambda), \qquad f_2(z) = \int_{-a}^{0} (z-\lambda)^{-1}\, d\gamma(\lambda).$$

Evidently, $f = f_1 + f_2$. We shall separately estimate the error of approximating $f_1(x)$ with an mth degree polynomial in x and $f_2(x)$ with an mth degree polynomial in x^{-1}. Let us denote by $c_k(g)$ the kth Fourier–Chebyshev coefficient of a function g on the interval $[\lambda_1, \lambda_n]$.

Begin with f_1. As $z \to -a$, $\Re z > -a$, we have, by virtue of (2.2),

$$|f_1(z)| \le \int_{-a-1}^{-a} |z-\lambda|^{-1}\, d\gamma(\lambda) + \int_{-\infty}^{-a-1} \left(1 + \frac{|z|}{|z-\lambda|}\right) \left|\lambda^{-1}\right|\, d\gamma(\lambda) \tag{3.2}$$

$$\le |z+a|^{-1} \gamma[-a-1, -a] + (1+|z|) \int_{-\infty}^{-a-1} \left|\lambda^{-1}\right|\, d\gamma(\lambda) \lesssim |z+a|^{-1}.$$

Hence, the function $(z+a)^2 f_1(z)$ is continuous on the ellipse in the complex plane with foci λ_1, λ_n, containing the point $-a$, and analytic inside the ellipse. Theorem 8.13 in [21][3] yields the estimate

$$c_k\left[(z+a)^2 f_1(z)\right] \lesssim \Phi(t)^{-k}. \tag{3.3}$$

Theorem 10.6 (formula (39)) in [21] gives the inequality

$$c_k\left[(z+a)^{-2}\right] \lesssim k\Phi(t)^{-k}. \tag{3.4}$$

[3]Let E_R be an ellipse with foci ± 1 and the sum of semi-axes R. If a function f is analytical in the open set, enclosed by the ellipse E_R, and is bounded on the ellipse so that

$$|f(z)| \le M_R,$$

then the Chebyshev coefficients $a_k(f)$ satisfy the inequality

$$|a_k(f)| \le 2M_R R^{-k}, \qquad k = 0, 1, \ldots.$$

Now, using Theorem 9.4 (formula (33)) in [21], devoted to the Chebyshev series of the product of two functions, we derive, from (3.3) and (3.4),

$$c_k(f_1) = c_k\left\{(z+a)^{-2}\left[(z+a)^2 f_1\right]\right. \quad \lesssim\ k^2\Phi(t)^{-k},$$

whence

$$\min_{p(X)\in\mathbf{R}[X],\ \deg p\le m} \|f_1 - p\|_{C[\lambda_1,\lambda_n]} \ \lesssim\ m^2\Phi(t)^{-m}. \tag{3.5}$$

Turn to f_2. Make the change of variables $y = a^2x^{-1},\ \ \psi = a^2\lambda^{-1}$, and define the measure μ on $[-\infty,-a]$ by

$$\mu(S) = \gamma\ \ a^2S^{-1}\big), \qquad S\subseteq[-\infty,-a] \text{ is an interval.}$$

We have

$$f_2(x) = f_2\ \ a^2y^{-1}\big) = a^{-2}y\int\limits_{-\infty}^{-a}(-\psi)(y-\psi)^{-1}\,d\mu(\psi),$$

and the fact that $\mu[-\infty,-a] = \gamma[-a,0] < +\infty$ implies that the function $f_2\ \ a^2z^{-1}\big)$ is analytical in $\mathbf{C}\backslash]-\infty,-a]$.

Note that $[a^2\lambda_n^{-1}, a^2\lambda_1^{-1}] = [\lambda_1,\lambda_n]$. As $z\to -a$, $\Re z > -a$, we get

$$\left|f_2\ \ a^2z^{-1}\big)\right| \ \lesssim\ \int\limits_{-\infty}^{-a}|\psi(z-\psi)^{-1}|\,d\mu(\psi)$$

$$\lesssim\ \int\limits_{-a-1}^{-a}|z-\psi|^{-1}\,d\mu(\psi) + \int\limits_{-\infty}^{-a-1}d\mu(\psi)\ \lesssim\ |z+a|^{-1}.$$

Analogously to the derivation of (3.5) from (3.2), this gives, for f_2,

$$\min_{q(Y)\in\mathbf{R}[Y],\ \deg q\le m}\ f_2\ \ a^2Y^{-1}\big) - q(Y)\ _{C[\lambda_1,\lambda_n]}\ \lesssim\ m^2\Phi(t)^{-m}. \tag{3.6}$$

Combining (3.5) and (3.6), we obtain (3.1). □

THEOREM 4. *Estimate* (3.1) *is also valid for any function from class* $\mathcal{B}$.

Proof. We use the induction by the minimal length of constructing a function f, belonging to class $\mathcal{B}$, according to the definition of class $\mathcal{B}$.

If f belongs to class $\mathcal{A}$, then f satisfies (3.1) due to Theorem 1.

If f is a linear combination of functions f_1 and f_2, both satisfying (3.1), then the linear combination of rational functions, providing the good estimate for f_1 and f_2, gives the desirable estimate for f.

Finally, suppose that $l\in\mathbf{Z}$ and

$$f(X) - X^{-m}p_m(X)\ _{C[\lambda_1,\lambda_n]}\ \lesssim\ m^2\Phi\,(t)^{-m}\,, \tag{3.7}$$

with t the same as in the proof of Theorem 1, $p_m \in \mathbf{R}[X]$, and $\deg p_m \le 2m$. If $l \ge 0$, we have, for $m > 2l$,

$$\begin{aligned}\min_{q(X)\in\mathbf{R}[X],\ \deg q\le 2m} & \left\| X^l f(X) - X^{-m} q(X) \right\|_{C[\lambda_1,\lambda_n]} \\ &\le \left\| X^l f(X) - X^{-m}[X^{3l} p_{m-2l}(X)] \right\|_{C[\lambda_1,\lambda_n]} \\ &\le \lambda_n^l \left\| f(X) - X^{-(m-2l)} p_{m-2l}(X) \right\|_{C[\lambda_1,\lambda_n]} \\ &\lesssim (m-2l)^2 \Phi(t)^{-m+2l} \lesssim m^2 \Phi(t)^{-m}\end{aligned}$$

owing to (3.7) and the inequality $\deg[X^{3l} p_{m-2l}(X)] \le 3l + 2(m-2l) \le 2m$. Analogously, in the case $l < 0$ we obtain, for $m > -l$,

$$\begin{aligned}\min_{q(X)\in\mathbf{R}[X],\ \deg q\le 2m} & \left\| X^l f(X) - X^{-m} q(X) \right\|_{C[\lambda_1,\lambda_n]} \\ &\le \left\| X^l f(X) - X^{-m}[p_{m+l}(X)] \right\|_{C[\lambda_1,\lambda_n]} \\ &\le \lambda_1^l \left\| f(X) - X^{-m-l}[p_{m+l}(X)] \right\|_{C[\lambda_1,\lambda_n]} \\ &\lesssim (m+l)^2 \Phi(t)^{-m-l} \lesssim m^2 \Phi(t)^{-m}.\end{aligned}$$

Anyway, estimate (3.1) holds. ∎

4. EKSM solution and an EKSM error estimate. Let W be the $n \times (k + m - 1)$ matrix whose columns form an orthonormal basis of $\mathcal{K}^{k,m}$, and let $\varphi = Ws$ with $s \in \mathbf{R}^{k+m-1}$. Introduce the $(k + m - 1) \times (k + m - 1)$ Ritz matrix R for A and W: $R = W^* A W$. We shall define the EKSM's approximant for (1.1) from $\mathcal{K}^{k,m}$ as

$$u_{k,m} = W f(R) s. \tag{4.1}$$

It is easily seen that, in exact arithmetic, vector (4.1) is independent of a particular choice of the basis W.

Now, we shall investigate the quality of approximate solution (4.1).

LEMMA 5. *If $g(X) = X^{-k+1} p(X)$, where $p(X) \in \mathbf{R}[X]$, $\deg p \le k + m - 2$, then*

$$g(A)\varphi = W g(R) s,$$

i.e., EKSM is exact for such a rational function g.

Proof. We can reckon that the columns of W are the $k + m - 1$ Lanczos vectors $w_1, w_2, \ldots$ of the Lanczos process with the matrix A and the vector $A^{-k+1}\varphi$. With this interpretation, we have

$$\varphi = \begin{array}{cc} A^{-k+1}\varphi & A^{k-1} w_1 \end{array}. \tag{4.2}$$

We shall twice use the exactness of the $k + m - 1$ steps of SLDM with respect to a polynomial of degree $\le k + m - 2$ (see [26]).

It follows from the definition of s and (4.2) that

$$\begin{aligned} s = W^*\varphi = & \begin{array}{cc} A^{-k+1}\varphi & W^* A^{k-1} w_1 \end{array} \\ = & \begin{array}{cc} A^{-k+1}\varphi & W^* W R^{k-1} e_1^{(k+m-1)} \end{array} = \begin{array}{cc} A^{-k+1}\varphi & R^{k-1} e_1^{(k+m-1)} \end{array}. \end{aligned} \tag{4.3}$$

Finally, obtain, with use of (4.3),

$$g(A)\varphi = p(A)\ \ A^{-k+1}\varphi\big) = \ \ A^{-k+1}\varphi\ \ Wp(R)e_1^{(k+m-1)}$$
$$= \ \ A^{-k+1}\varphi\ \ Wg(R)R^{k-1}e_1^{(k+m-1)} = Wg(R)s. \qquad \square$$

THEOREM 6. *Let a function f belong to class $\mathcal{B}$, and let $k = m$. Then the error estimate*

$$\|f(A)\varphi - Wf(R)s\| \ \lesssim\ m^2\Phi\left(\frac{1+\sqrt{\frac{\lambda_1}{\lambda_n}}}{1-\sqrt{\frac{\lambda_1}{\lambda_n}}}\right)^{-m}$$

holds for "diagonal" approximants in EKSM.

Proof. Put $g(X) = X^{-m+1}p(X)$ with $p(X) \in \mathbf{R}[X]$, $\deg p \le 2m-2$. Using Lemma 1 and the fact that the Ritz values lie on $[\lambda_1, \lambda_n]$, we deduce

$$\|f(A)\varphi - Wf(R)s\| \le \|(f-g)(A)\varphi\| + \|W(f-g)(R)s\| \ \lesssim\ \|f-g\|_{C[\lambda_1,\lambda_n]}.$$

Owing to Theorem 2, it only remains to use (3.1). $\square$

For small positive x we can use the estimate $\Phi\,(1+x)^{-m} \approx \exp\ \ -m\sqrt{2x}\big)$, so for large values of the condition number λ_n/λ_1, Theorem 3 yields

$$\|f(A)\varphi - Wf(R)s\| \cong O\left[\exp\left(-2m\sqrt[4]{\frac{\lambda_1}{\lambda_n}}\right)\right]. \tag{4.4}$$

For comparison, for a large condition number approximation on $\mathcal{K}^J(A,\varphi)$ (SLDM) [4, Theorem 3] converges as

$$\cong O\left[\Phi\left(\frac{1+\frac{\lambda_1}{\lambda_n}}{1-\frac{\lambda_1}{\lambda_n}}\right)^{-J}\right] \cong O\left[\exp\left(-2J\sqrt{\frac{\lambda_1}{\lambda_n}}\right)\right]. \tag{4.5}$$

This estimate can be easily extended to approximations on $\mathcal{K}^J(A^{-1},\varphi)$. Comparing (4.4) and (4.5), we conclude that, to get a fixed approximation quality in EKSM and SLDM, one has to take $m \asymp \sqrt{J}$.

5. The extended Lanczos recurrence. First, we shall derive an economical procedure for computing an orthonormal basis of $\mathcal{K}^{k,m}$.

1°. Let us perform the first k steps of the Lanczos recurrence with $B = A^{-1}$ and φ [20]. We shall denote by Q the obtained $n \times k$ matrix of Lanczos vectors $Q = (q_1, \dots, q_k)$ and by H the $k \times k$ tridiagonal symmetric Ritz matrix

$$H = \begin{pmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \\ & & \ddots & \\ 0 & & \beta_{k-1} & \alpha_k \end{pmatrix}.$$

They are related by the standard matrix formulae [20, section 13.1]

$$BQ - QH = re_k^{(k)*}, \quad Q^*Q = I, \quad r = \beta_k q_{k+1}. \tag{5.1}$$

2°. Let us obtain the vector v_1 from the formula

$$b_0^{(1)} v_1 = Aq_1 - \sum_{i=1}^{k} c_i q_i, \tag{5.2}$$

where the coefficients c_i are selected to make v_1 orthogonal to q_i, $i = 1, \ldots, k$, and $b_0^{(1)} > 0$ is determined by the condition $\|v_1\| = 1$. Analogously, obtain the vector v_2 from

$$b_1 v_2 = Av_1 - a_1 v_1 - \sum_{i=1}^{k} b_0^{(i)} q_i, \tag{5.3}$$

where the coefficients $b_0^{(i)}$ and a_1 are selected to make v_2 orthogonal to q_i, $i = 1, \ldots, k$, and v_1, respectively. Formulae (5.2) and (5.3) are special cases of the rational Krylov subspaces (RKS) recurrence ([25, p. 286]), which relates all of the previously computed orthonormal vectors. However, we will show how to realize (5.2) and (5.3) without keeping the whole matrix Q in memory.

3°. Suppose that we have constructed an orthonormal basis $(q_1, \ldots, q_k, v_1, \ldots, v_i)$ for $\mathcal{K}^{k,i+1}$, $i \geq 2$. If $w \in \mathcal{K}^{k,i-1}$, then $\langle Av_i, w\rangle = \langle v_i, Aw\rangle = 0$, because $Aw \in \mathcal{K}^{k-1,i} \subseteq \mathcal{K}^{k,i}$ and $v_i \perp \mathcal{K}^{k,i}$. So, we have arrived at a three-term Gram–Schmidt recurrence similar to the standard Lanczos algorithm

$$b_i v_{i+1} = Av_i - a_i v_i - b_{i-1} v_{i-1}, \qquad 2 \leq i \leq m-1. \tag{5.4}$$

The desired orthonormal basis of $\mathcal{K}^{k,m}$ has been constructed.

Define the $n \times (k+m-1)$ matrix $W = (q_1 \ldots q_k \, v_1 \ldots v_{m-1})$. By construction, $W^*W = I$. Let $R = W^*AW$ be the Ritz matrix. In block form,

$$R = \begin{pmatrix} G & D \\ D^* & T \end{pmatrix},$$

where G is, generally speaking, a full $k \times k$ matrix, and it can be seen from (5.3)–(5.4) that T is an $(m-1) \times (m-1)$ tridiagonal symmetric matrix

$$T = \begin{pmatrix} a_1 & b_1 & & 0 \\ b_1 & a_2 & b_2 & \\ & & \ddots & \\ 0 & & b_{m-2} & a_{m-1} \end{pmatrix},$$

and D is an $(m-1) \times k$ matrix of rank 1 with only the first nonzero column

$$D = \left(b_0^{(1)} \ldots b_0^{(k)}\right)^* e_1^{(m-1)*}.$$

Now we shall deduce an expression for the submatrix $G = Q^*AQ$. In a term e_i, the superscript (k) will be assumed on default.

From (5.1) we have

$$AQ = QH^{-1} - Are_k^* H^{-1} \tag{5.5}$$

and

$$G = Q^*AQ = H^{-1} - Q^*Are_k^* H^{-1} = H^{-1} - (AQ)^* re_k^* H^{-1}. \tag{5.6}$$

Further, substituting (5.5) in (5.6) and using orthogonality of r and q_i, $i = 1, \dots, m$, we get

$$G = H^{-1} - \ QH^{-1}\big)^* re_k^* H^{-1} + \ Are_k^* H^{-1}\big)^* re_k^* H^{-1} \tag{5.7}$$
$$= H^{-1} + \ Are_k^* H^{-1}\big)^* re_k^* H^{-1} = H^{-1} + (r^* Ar) \ H^{-1}e_k\big) \ H^{-1}e_k\big)^*.$$

So, we have arrived at a formula for G that requires computing only one inner product r^*Ar.

We promised to rewrite (5.2)–(5.3) in a form more convenient for computation. With use of (5.5) and (5.7), deduce

$$b_0^{(1)} v_1 = AQe_1 - QGe_1 = QH^{-1}e_1 - Are_k^* H^{-1}e_1 - QGe_1 \tag{5.8}$$
$$= QH^{-1}e_1 - Are_k^* H^{-1}e_1 - Q\left[H^{-1} + (r^*Ar) \ H^{-1}e_k\big) \ H^{-1}e_k\big)^*\right] e_1$$
$$= -e_k^* H^{-1} e_1 \left[Ar + (r^*Ar)\, QH^{-1}e_k\right].$$

Now, derive an efficient formula for $b_0^{(i)}$. By definition, and by means of (5.5),

$$b_0^{(i)} = v_1^* Aq_i = v_1^* AQe_i = v_1^* QH^{-1}e_i - v_1^* Are_k^* H^{-1}e_i = -v_1^* Are_k^* H^{-1}e_i.$$

Since (5.2) and (5.8) give

$$b_0^{(1)} = \left\langle b_0^{(1)} v_1, v_1 \right\rangle = - \ e_k^* H^{-1} e_1\big)\, v_1^* \left[Ar + (r^*Ar) QH^{-1}e_k\right] = - \ e_k^* H^{-1} e_1\big)\, (v_1^* Ar),$$

we get

$$b_0^{(i)} = \frac{b_0^{(1)}}{e_k^* H^{-1} e_1} e_k^* H^{-1} e_i, \qquad i = 2, \dots, k. \tag{5.9}$$

Expression (5.9) allows us to rewrite (5.3) in the fashion of (5.8),

$$b_1 v_2 = Av_1 - a_1 v_1 - \frac{b_0^{(1)}}{e_k^* H^{-1} e_1} QH^{-1}e_k. \tag{5.10}$$

It is very convenient that the same vector $QH^{-1}e_k$ is used in both formula (5.8) and (5.10).

4°. Computing $QH^{-1}e_k$ for (5.8) and (5.10) can be done recursively in the CG fashion, without keeping the whole matrix Q in the memory. Indeed, denote by H_i the left upper $i \times i$ submatrix of H, $X_i = (x_{1i}, x_{2i}, \dots, x_{ii})^* = H_i^{-1} e_i^{(i)}$, $Q_i = (q_1 \dots q_i)$, $p_i = x_{ii}^{-1} Q_i X_i$, and finally $QH^{-1}e_k = x_{kk} p_k$.

Using the well known connection [12, subsect. 9.3.1] between the Lanczos and CG methods, we can state that p_i is the ith B-orthogonal vector used in the CG recurrence. We derive a simple CG-like recurrence for p_i.

It can be checked by direct substitution that, for $i > 1$,

$$H_{i+1}\left[-\sigma_i s^{(i+1)} + e_{i+1}^{(i+1)}\right] = x_{i+1,i+1}^{-1} e_{i+1}^{(i+1)}, \tag{5.11}$$

where $s^{(i+1)} = x_{ii}^{-1}(x_{1i}, x_{2i}, \dots, x_{ii}, 0)^*$, $\sigma_i = -x_{i+1,i+1}^{-1} x_{i,i+1}$. Equality (5.11) evidently yields $-\sigma_i s^{(i+1)} + e_{i+1}^{(i+1)} = x_{i+1,i+1}^{-1} X_{i+1}$. Then, using the formulae $p_1 = q_1$, $p_i = Q_{i+1} s^{(i+1)}$, and $q_{i+1} = Q_{i+1} e_{i+1}^{(i+1)}$, we obtain the recurrence

$$p_{i+1} = -\sigma_i p_i + q_{i+1}.$$

We can define σ_i recursively as well. Rewriting the ith equation of the system $H_{i+1}X_{i+1} = e_{i+1}^{(i+1)}$ componentwise, we obtain

$$\beta_{i-1}x_{i-1,i+1} + \alpha_i x_{i,i+1} + \beta_i x_{i+1,i+1} = 0. \tag{5.12}$$

By definition, $x_{i,i+1} = -\sigma_i x_{i+1,i+1}$, and (5.11) yields $x_{i-1,i+1} = -\sigma_{i-1}x_{i,i+1}$, so we can rewrite (5.12) as

$$\beta_{i-1}\sigma_{i-1}\sigma_i - \alpha_i\sigma_i + \beta_i = 0. \tag{5.13}$$

Finally, we can write the recurrence starting from the vector $p_1 = q_1$ and the scalar $\sigma_0 = 0$ as

$$\sigma_i = \frac{\beta_i}{\alpha_i - \beta_{i-1}\sigma_{i-1}}, \quad p_{i+1} = -\sigma_i p_i + q_{i+1}, \quad 1 \le i \le k-1. \tag{5.14}$$

Performing (5.14) simultaneously with the first Lanczos recurrence, we do not need to keep the additional vectors p_i in memory.

Given k and m, the algorithm just described can be summarized as follows.

1. Perform the k steps of the Lanczos method with A^{-1} and φ. Obtain Q, H, and r. At the same time, perform recurrence (5.14), producing p_k.
2. Compute r^*Ar.
3. Invert H.
4. Compute $QH^{-1}e_k = \left(e_k^*H^{-1}e_k\right)p_k$.
5. Compute $b_0^{(1)}$ and v_1 by means of (5.8) and further normalization.
6. Compute $b_0^{(2)}, \ldots, b_0^{(k)}$ by means of (5.9).
7. Compute v_2, a_1, and b_1 by means of (5.10) and further normalization.
8. Perform recurrence (5.4). Obtain T.
9. Complete the computation of R, calculating G by means of (5.7).

6. The EKSM matrix-functional formula. Let W and R be defined as in section 5. Since $\varphi = \|\varphi\|We_1^{(k+m-1)}$, we can rewrite formula (4.1) in the SLDM fashion as

$$u_{k,m} = \|\varphi\|Wf(R)e_1^{(k+m-1)}. \tag{6.1}$$

The clearest way to compute

$$\dot{f}^{(k,m)} = f(R)e_1^{(k+m-1)},$$

exploiting general symmetric eigensolvers, is to use the eigendecomposition of R. However, the structure of R allows one to factorize it cheaply and to use rational approximations.

Straightforward application of EKSM requires storing the whole matrix W; that may be quite costly. However, for some applications it is not necessary to compute all the n components of the vector $u_{k,m}$, but only a few of them. Then only W's rows corresponding to these components have to be stored. Another option would be a two stage algorithm: first to perform the Extended Lanczos Recurrence to compute only $\dot{f}^{(k,m)}$ storing the recurrence coefficients and only vectors q and v currently necessary for the recurrence, then to recompute q and v using the recurrence coefficients obtained before calculating synchronously with $W\dot{f}^{(k,m)}$.

Let $u_{k,m}$ be the EKSM approximation to (2.9). Evidently, $u_{k,m}$ satisfies the boundary conditions, so the residual of $u_{k,m}$ in (1.2) can be theoretically used as an a posteriori estimate for $\|u_{k,m} - w\|$. The extended Lanczos recurrence can be written in the matrix form, similar to the standard Lanczos recurrence (5.1), as

$$AW - WR = b_{m-1} v_m e_{k+m-1}^{(k+m-1)*}, \quad W^*W = I, \quad m \geq 2, \tag{6.2}$$

which can be derived from the definitions of W and R and from (5.4) with $i = m-1$. Combining (6.2) and (6.1), we obtain a standard expression for the residual, similar to the well-known formula for CG,

$$Au_{k,m} - \frac{d^2 u_{k,m}}{d\theta^2} = \|\varphi\| b_{m-1} \left(e_{k+m-1}^{(k+m-1)*} \dot{f}^{(k,m)} \right) v_m, \tag{6.3}$$

that is valid for the problem outlined in Example 3 as well.

7. Solution of an elliptic problem by the method of lines. We consider the 2.5-D direct current problem in the cylindrical coordinate system (r, z, θ). Let $\sigma(r, z)$ be axially symmetric conductivity and let $U(r, z, \theta)$ be the potential of a source $\delta(\theta) r^{-1} g(r, z)$. This problem arises from geophysical computed tomography, when axially symmetric conductivity perturbation is mapped due to direct current (DC) injections and measurements from asymmetrically placed wells. It is crucial to have a fast forward solver for the inversion.

We state the PDE

$$r^{-1} \frac{\partial}{\partial r} \left(r\sigma \frac{\partial U}{\partial r} \right) + \frac{\partial}{\partial z} \left(\sigma \frac{\partial U}{\partial z} \right) + r^{-2} \sigma \frac{\partial^2 U}{\partial \theta^2} = - \sum_{k=-\infty}^{\infty} \delta(\theta - 2\pi k) \delta(r - r_0) \tag{7.1}$$

in a bounded rotational domain $\subset \mathbf{R}^3$; the trivial Dirichlet boundary condition at ∂ is assumed. Substituting in (7.1) $u = U\sqrt{r^{-1}\sigma}$, obtain

$$Au - \frac{\partial^2 u}{\partial \theta^2} = \sum_{k=-\infty}^{\infty} \delta(\theta - 2\pi k) \varphi, \tag{7.2}$$

where A is a symmetric positive definite operator

$$Au = -\sqrt{\frac{r}{\sigma}} \frac{\partial}{\partial r} \left[r\sigma \frac{\partial}{\partial r} \left(\sqrt{\frac{r}{\sigma}}\, u \right) \right] - \frac{r^2}{\sqrt{\sigma}} \frac{\partial}{\partial z} \left[\sigma \frac{\partial}{\partial z} \left(\frac{1}{\sqrt{\sigma}}\, u \right) \right] \tag{7.3}$$

and

$$\varphi = \sqrt{\frac{r^3}{\sigma}}\, \delta(r - r_0).$$

We approximate (7.1) on a set of axial circular lines, introducing a second order finite differences (FD) discretization of operator (7.3) on a 2-D five-point grid with n nodes. Then equation (7.2) becomes an ODE system with respect to the n-dimensional vector-function $u(\theta)$.

Using (1.2), we get a matrix functional representation of u

$$u(\theta) = 0.5 \left[e^{-\theta\sqrt{A}} + e^{-(2\pi-\theta)\sqrt{A}} \right] \left[\sqrt{A} \left(I - e^{-2\pi\sqrt{A}} \right) \right]^{-1} \varphi, \qquad 0 \leq \theta \leq \pi. \tag{7.4}$$

We applied EKSM to matrix function (7.4). For computing the LU decomposition of A^{-1} and its action, we used the subroutine MA27D from HARWELL package. LAPACK's spectral solver DSBEV was adopted for the full eigenproblem of R.

For example, we took a resistive disk in a homogeneous medium (piston-like invasion of fresh water from the well): the point source at $(1000, 0, 0)$ and the point receiver at $(1000, -250, 0)$. The grid contained $n = 4400$ nodes, $\lambda_1 = .3$, $\lambda_n = 9 \cdot 10^6$.

In Fig. 1, we plotted the relative errors produced by the standard SLDM versus m at the receiver location together[4] with a priori estimate (4.5). Actually, due to computer roundoff, m may exceed n. However, our analysis [5, 6] of SLDM's stability, based on Paige's results on the simple Lanczos recurrence without reorthogonalization, [18] proves that the exact arithmetic estimate (4.5) is still valid in the computer. This phenomenon is clearly seen in the figure.

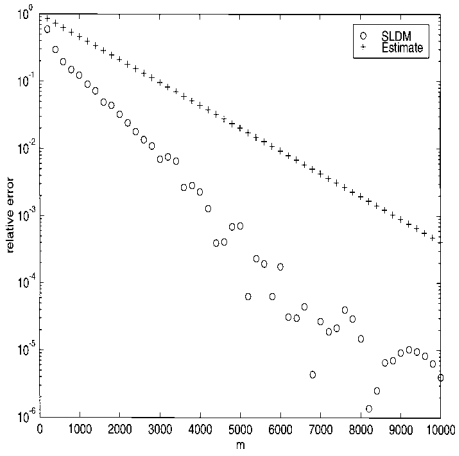


FIG. 1. *Actual SLDM convergence and a priori estimate.*

To converge in five decimal digits, SLDM took 9300 Lanczos steps and 29 seconds on IBM Risk-6000. Approximation on $\mathcal{K}^m(A^{-1}, \varphi)$ exhibits similar behavior but requires larger CPU time.

Basically, we observed the same behavior in the extended Lanczos recurrence as in the simple Lanczos recursion in the computer [11]: loss of orthogonality of W and arriving spurious Ritz values, clustered in tiny vicinities of the well separated eigenvalues of A, mainly at the bounds of the spectral interval. We did not make any computer arithmetic analysis of EKSM similar to [5] because there are not extensions of [18] in the literature. However, our experiments showed that, similar to SLDM, the observed instability of the extended Lanczos recurrence does not affect the quality of the converged EKSM results: the process is unstable itself, but the bound of Theorem 3 is valid in the computer up to roundoff.

In Fig. 2 we plotted the EKSM's relative error versus m at the receiver location for the diagonal subspaces $\mathcal{K}^{m,m}$, together with the corresponding estimate of Theorem 3. To compute the relative error, we consider the solution of the given finite-difference problem, obtained for 500 steps of EKSM, as exact. Though the actual EKSM error exhibits some irregularities, it converges much faster than the estimate. A similar

[4]Drawing the graphs of estimates, we threw away indefinite coefficients hidden in O symbols.

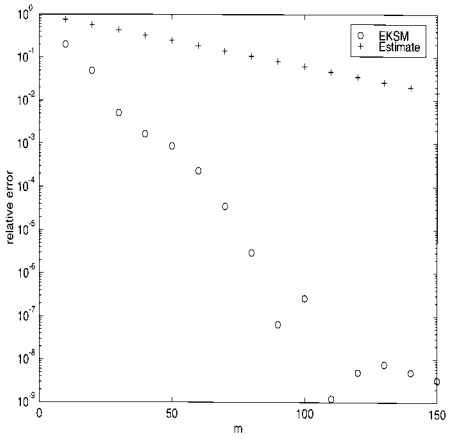


FIG. 2. *Actual EKSM convergence for $k = m$ and estimate of Theorem* 3.

phenomenon is normally observed in the behavior of actual CG errors compared to their Chebyshev bounds [11].

Normally we implement EKSM, first executing a prescribed number k of the negative steps, then performing the positive iterations until the results reach a suitable level of convergence. An example of such iterations with $k = 50$ is shown in Fig. 3. The standard EKSM without reorthogonalization (presented by the circles) converged to level 10^{-5} with $m = 130$, and it took 3.3 seconds, so we observed *one order reduction* in CPU time compared to SLDM.

In this example the solution is highly singular; that is typical for geophysical applications, i.e., the residual is useless for an a posteriori estimate of the relative error at the receivers because the required components of the solution can be extremely small compared to $\|u\|$. Instead, practitioners normally use empirical pointwise convergence criteria. Here, in qualitative (not quantitative) accordance with the formula from Theorem 3, we assume that the error at the selected receiver decreases as a geometrical series with an indefinite coefficient and multiplier. The result of three steps enables us to predict the error by solving a simple system of algebraic equations. This procedure is performed not too often in order to avoid false apparent "convergence," i.e., typically we select a subsequence of control steps $l, 2l, 3l, \ldots$ with a constant increment l; here we selected $l = 15$. The asterisks in Fig. 3 present our sample convergence criterion; it produces reasonable estimates of the order of the relative error.

The Rational Krylov Subspaces algorithm, using zero and infinite shifts, is equivalent to the EKSM with complete reorthogonalization. In exact arithmetic the short EKSM recursion is certainly better than full Gram–Schmidt orthogonalization. However, in the computer the EKSM behaves similar to the Lanczos method, i.e., loss of orthogonality sometimes slows down the convergence.

The third graph in Fig. 3 presents the results of an experiment performed with full reorthogonalization on $\mathcal{K}^{50,m}$. The reorthogonalization showed error reduction up to one order at the expense of a significant increase in computational cost. The reorthogonalized algorithm converged to five digits with $m = 80$ and it took 11 seconds (compared to 3.3 seconds for the short recursion producing the same accuracy), so

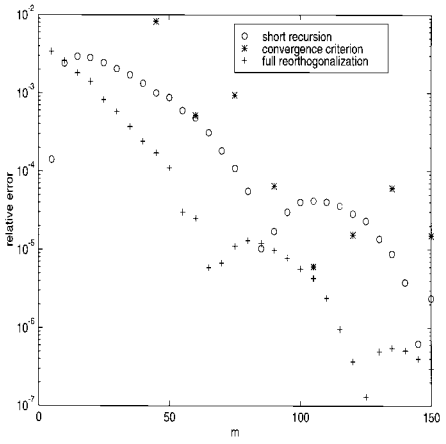


FIG. 3. *Convergence of the EKSM using the short recurrence and full reorthogonalization.*

additional expenses on reorthogonalization by far overwhelmed the observed decrease of m. These expenses grow as $O[n(k+m)^2]$, so they would affect computational performance even more strongly if one wants to achieve a higher accuracy! In addition, the whole matrix W has to be stored in order to perform the reorthogonalization; in our example with reorthogonalization, the capability of RAM (128 MB) was exceeded when m reached 150, afterwards the computational performance was affected enormously due to use of the virtual memory.

8. Additional remarks. It looks attractive to apply EKSM to indefinite elliptic problems (waveguides) and stable parabolic problems (exponential propagation). Another feasible application of EKSM can be a linear system with a matrix, which can be expressed as the sum of two easily invertible matrices.

Acknowledgments. The authors thank Sheng Fang, Tarek Habashy, Shari Moskow, and Carlos Torres-Verdìn for useful discussions, Paul Van Dooren for supplying a preliminary copy of [9], and Hans Blok and Mathe van Stralen for providing some references.

REFERENCES

[1] A. ALLERS, A. SEZGINER, AND V. DRUSKIN, *Solution of* 2.5*-dimensional problems using the Lanczos decomposition*, Radio Sci., 29 (1994), pp. 955–963.
[2] G. BAKER AND P. GRAVES-MORRIS, *Padé Approximants*, Addison–Wesley, London, 1981.
[3] V. L. DRUSKIN AND L. A. KNIZHNERMAN, *A spectral semi-discrete method for the numerical solution of* 3*D nonstationary problems in electrical prospecting*, Izv. Akad. Sci. U.S.S.R., Physics of Solid Earth, 24 (1988), pp. 641–648. (English edition published by American Geophysical Union.)
[4] V. L. DRUSKIN AND L. A. KNIZHNERMAN, *Two polynomial methods of calculating functions of symmetric matrices*, USSR Comput. Math. Math. Phys., 29 (1989), pp. 112–121.
[5] V. L. DRUSKIN AND L. A. KNIZHNERMAN, *Error bounds in the simple Lanczos procedure for computing functions of symmetric matrices and eigenvalues*, USSR Comput. Math. Math. Phys., 31 (1991), pp. 20–30.
[6] V. L. DRUSKIN, A. GREENBAUM, AND L. A. KNIZHNERMAN, *Using nonorthogonal Lanczos vectors in the computation of matrix functions*, SIAM J. Numer. Anal., 19 (1998), to appear.

[7] I. S. Duff, A. M. Erisman, and J. K. Reid, *On George's nested dissection method*, SIAM J. Numer. Anal., 13 (1976), pp. 686–695.
[8] R. A. Frishner, L. S. Tukerman, B. C. Dornblacer, and T. V. Russo, *A method of exponential propagation of large systems of stiff nonlinear differential equations*, J. Sci. Comput., 4 (1989), pp. 327–335.
[9] K. Gallivan, E. Grimme, D. Sorensen, and P. Van Dooren, *On some modification of the Lanczos algorithm and the relation with Padé approximations*, in Proc. ICIAM '95, Hamburg, Math. Res. 87, Akademie Verlag, Berlin, 1996, pp. 87–116.
[10] E. Gallopoulos and Y. Saad, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1236–1264.
[11] A. Greenbaum, *Behavior of slightly perturbed Lanczos and conjugate-gradient recurrences*, Linear Algebra Appl., 113 (1989), pp. 7–63.
[12] G. Golub and Ch. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1989.
[13] B. Hermansson, D. Yevick, W. Bardyszewski, and M. Glasner, *A comparison of Lanczos electric field propagation methods*, IEEE J. Light Wave Technology, 10 (1992), pp. 772–776.
[14] P.-L. Liu and B.-J. Li, *Semivectorial Helmholtz beam propagation by Lanczos reduction*, IEEE J. Quantum Electronix, 29 (1993), pp. 2385–2389.
[15] A. Nauts and R. E. Wyatt, *New approach to many state quantum dynamics*, Phys. Rev. Lett., 51 (1983), pp. 2238–2241.
[16] E. M. Nikishin and V. N. Sorokin, *Rational Approximations and Orthogonality*, Nauka, Moscow, 1988 (in Russian).
[17] B. Nour-Omid and R. W. Clough, *Dynamic analysis of structure using Lanczos co-ordinates*, Earthquake Engrg. Structur. Dynamics, 12 (1984), pp. 565–577.
[18] C. C. Paige, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, Ph.D. thesis, University of London, London, 1971.
[19] T. J. Park and J. C. Light, *Unitary quantum time evolution by iterative Lanczos reduction*, J. Chem. Phys., 85 (1986), pp. 5870–5876.
[20] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[21] S. Paszkowski, *Computational Applications of Chebyshev Polynomials and Series*, Nauka, Moscow, 1983 (in Russian; translation from Polish).
[22] R. Ratowsky, J. Fleck, Jr., and M. Feit, *Helmholtz beam propagation in rib waveguides and couplers by iterative Lanczos reduction*, J. Optical Soc. Amer., 9 (1992), pp. 265–273.
[23] A. Ruhe, *Rational Krylov sequence methods for eigenvalue computation*, Linear Algebra Appl., 58 (1984), pp. 391–405.
[24] A. Ruhe, *Rational Krylov algorithms for nonsymmetric eigenvalue problems*, in Recent Advances in Iterative Methods, IMA Volumes in Mathematics and its Applications 60, G. Golub, A. Greenbaum, and M. Luskin, eds., Springer-Verlag, Berlin, 1993, pp. 149–164.
[25] A. Ruhe, *Rational Krylov algorithms for nonsymmetric eigenvalue problems,* II*: Matrix pairs*, Linear Algebra Appl., 197/198 (1994), pp. 283–295.
[26] Y. Saad, *Projection methods for solving large sparse eigenvalue problems*, in Matrix Pencils, Proceedings, Pitea Havsbad, Springer-Verlag, Berlin, 1982, pp. 121–144.
[27] T. Tamarchenko, *Fast Algorithms of Electromagnetic Forward Modeling in Complex Geometry*, Ph.D. thesis, Moscow Geol. Prospect. Inst., Moscow, 1988 (in Russian).
[28] H. A. van der Vorst, *An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A*, J. Comput. Appl. Math., 18 (1987), pp. 249–263.
[29] W. Weedon, W. Chew, J-H. Lin, A. Sezginer, and V. Druskin, *A 2.5-D scalar Helmholtz wave solution employing the Spectral Lanczos Decomposition Method (SLDM)*, Microwave Optical Tech Lett., 6 (1993), pp. 587–592.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 772–775, July 1998

011

INVERSE ORDER RULE FOR WEIGHTED GENERALIZED INVERSE*

WENYU SUN† AND YIMIN WEI‡

Abstract. The weighted generalized inverses have several important applications in researching the singular matrices, regularization methods for ill-posed problems, optimization problems, and statistics problems. In this paper we establish some sufficient and necessary conditions for inverse order rule of weighted generalized inverse.

Key words. generalized inverse, weighted generalized inverse, inverse order rule, matrix computation

AMS subject classifications. 15A09, 65F10

PII. S0895479896305441

1. Introduction.

The generalized inverse is an important tool for researching the singular matrix problems, ill-posed problems, optimization problems, and statistics problems. The inverse order rule for generalized inverse plays an important role in the theoretic research and numerical computations in the above areas (see [2], [4], [5], [6], [8], [11], [12], [13], [14], [15]). The purpose of this paper is to establish the inverse order rule of the weighted generalized inverse, which is met in our research of the weighted trust region approach for optimization problems [3], [7], [9], [10]. In addition, the inverse order rule for weighted generalized inverse is also applied to the generalized least squares problem and the weighted perturbation theory of the singular matrix.

In general, the inverse order rule does not always hold. If $A \in C^{m\times n}, B \in C^{n\times l}$, the sufficient and necessary condition for inverse order rule is as follows.

LEMMA 1.1. *The following conditions are equivalent:*

1. $(AB)^+ = B^+A^+$;
2. $\mathcal{R}(A^*AB) \subset \mathcal{R}(B)$ *and* $\mathcal{R}(BB^*A^*) \subset \mathcal{R}(A^*)$;
3. $\mathcal{R}(A^*ABB^*) = \mathcal{R}(BB^*A^*A)$.

(*See* [1], [2], [5], [6].)

In this paper we generalize the above results to the case of the weighted generalized inverse.

2. Main results.

Notation. For convenience, we list some notation as follows:

$C^{m\times n}, C_r^{m\times n}$: $m\times n$ matrix set and $m\times n$ matrix set with rank r, respectively;
$\mathcal{R}(\cdot), \mathcal{N}(\cdot)$: range and null space, respectively;
A^*: conjugate transpose matrix of A;
$A^\#$: weighted conjugate transpose matrix of A;
A^+: Moore–Penrose inverse of A;
$A^+_{M,N}$: weighted Moore–Penrose inverse of A;

* Received by the editors June 19, 1996; accepted for publication (in revised form) by G. P. Styan August 29, 1997; published electronically April 2, 1998. This work was supported by CNPq of Brazil and the National Natural Science Foundation of China.
http://www.siam.org/journals/simax/19-3/30544.html

† Department of Mathematics, Nanjing Normal University, Nanjing 210097, People's Republic of China (sun@mat.ufpr.br).

‡ Department of Mathematics, Fudan University, Shanghai 200433, People's Republic of China (ymwei@fudan.edu.cn).

$P_{\mathcal{L},\mathcal{M}}$: a projector with range $\mathcal{L}$ and null space $\mathcal{M}$.

DEFINITION 2.1. *Let $A \in C^{m\times n}$. Also, let M and N be $m\times m$ and $n\times n$ positive definite Hermite matrices, respectively. Then there is a unique matrix $G \in C^{n\times m}$ such that*

$$AGA = A, GAG = G, (MAG)^* = MAG, (NGA)^* = NGA, \tag{1}$$

where G is called weighted Moore–Penrose generalized inverse and written as $G = A^+_{MN}$.

DEFINITION 2.2. *Let M and N be $m \times m$ and $n \times n$ positive definite matrices, respectively. Given $A \in C^{m\times n}$, the weighted conjugate transpose matrix $A^\#$ of A is defined as*

$$A^\# = N^{-1}A^*M. \tag{2}$$

Obviously, $A^\#$ satisfies the following properties: if $A, A_1 \in C^{m\times n}, B \in C^{n\times l}$, then

$$(A + A_1)^\# = A^\# + A_1^\#, (AB)^\# = B^\#A^\#, (A^\#)^\# = A, (A^\#)^* = (A^*)^\#. \tag{3}$$

Before giving the properties of weighted generalized inverse, we state one lemma which is proved in [2] and [5].

LEMMA 2.3. *A^+_{MN} satisfies the following properties:*

1. $AA^+_{MN} = P_{\mathcal{R}(A),M\ {}^1\mathcal{N}(A^*)} = P_{R(A),N(A^\#)}$, $A^+_{MN}A = P_{N\ {}^1\mathcal{R}(A^*),\mathcal{N}(A)} = P_{R(A^\#),N(A)}$;
2. $A^+_{MN} = N^{-\frac{1}{2}}(M^{\frac{1}{2}}AN^{-\frac{1}{2}})^+M^{\frac{1}{2}}$.

In the following, we establish some sufficient and necessary conditions for the inverse order rule of the weighted generalized inverse. Here we employ a brief proof instead of the original proof due to a referee's suggestion.

THEOREM 2.4. *Let $A \in C^{m\times n}, B \in C^{n\times l}$. Also, let M, N, L be $m \times m, n \times n$, and $l \times l$ positive definite Hermite matrices, respectively. Then*

$$(AB)^+_{ML} = B^+_{NL}A^+_{MN} \tag{4}$$

if and only if

$$\mathcal{R}(A^\#AB) \subset \mathcal{R}(B) \text{ and } \mathcal{R}(BB^\#A^\#) \subset \mathcal{R}(A^\#). \tag{5}$$

Proof. In view of Lemma 2.3, we have

$$B^+_{NL}A^+_{MN} = (AB)^+_{ML} \tag{6}$$

if and only if

$$L^{-\frac{1}{2}}(N^{\frac{1}{2}}BL^{-\frac{1}{2}})^+N^{\frac{1}{2}}N^{-\frac{1}{2}}(M^{\frac{1}{2}}AN^{-\frac{1}{2}})^+M^{\frac{1}{2}} = L^{-\frac{1}{2}}(M^{\frac{1}{2}}ABL^{-\frac{1}{2}})^+M^{\frac{1}{2}},$$

or, equivalently, if and only if

$$\tilde{B}^+\tilde{A}^+ = (\tilde{A}\tilde{B})^+, \tag{7}$$

where

$$\tilde{A} := M^{\frac{1}{2}}AN^{-\frac{1}{2}} \text{ and } \tilde{B} := N^{\frac{1}{2}}BL^{-\frac{1}{2}}.$$

Lemma 1.1 tells us that (7) holds if and only if

$$\mathcal{R}(\tilde{A}^*\tilde{A}\tilde{B}) \subset \mathcal{R}(\tilde{B}) \text{ and } \mathcal{R}(\tilde{B}\tilde{B}^*\tilde{A}^*) \subset \mathcal{R}(\tilde{A}^*). \tag{8}$$

That (8) is equivalent to

$$\mathcal{R}(A^{\#}AB) \subset \mathcal{R}(B) \text{ and } \mathcal{R}(BB^{\#}A^{\#}) \subset \mathcal{R}(A^{\#}) \tag{9}$$

follows easily from the definition of $(\cdot)^{\#}$ and the fact that M, N, and L are positive definite matrices. This completes the proof. □

COROLLARY 2.5. *Let $A \in C^{m\times n}, B \in C^{n\times l}$. Also, let M, N, L be $m \times m, n \times n$, and $l \times l$ positive definite Hermite matrices, respectively. Then*

$$(AB)^{+}_{ML} = B^{+}_{NL}A^{+}_{MN}$$

if and only if

$$A^{+}_{MN}ABB^{\#}A^{\#} = BB^{\#}A^{\#} \text{ and } BB^{+}_{NL}A^{\#}AB = A^{\#}AB. \tag{10}$$

Proof. It is directly obtained from (5) and Lemma 2.3. □

THEOREM 2.6. *Let $A \in C^{m\times n}, B \in C^{n\times l}$. Also, let M, N, L be $m \times m, n \times n$, and $l \times l$ positive definite Hermite matrices. Then*

$$(AB)^{+}_{ML} = B^{+}_{NL}A^{+}_{MN} \tag{11}$$

if and only if

$$\mathcal{R}(A^{\#}ABB^{\#}) = \mathcal{R}(BB^{\#}A^{\#}A). \tag{12}$$

Proof. Similar to the proof of Theorem 2.4, we can also directly obtain this result from Lemmas 1.1 and 2.3. □

Acknowledgments. The authors would like to thank the referees and the associate editor George P. H. Styan for their helpful suggestions. The authors would especially like to express their gratitude to Dr. Hans Joachim Werner for providing a brief proof for their main result instead of the original version of the proof and two important references [13], [14] which improved the paper greatly.

REFERENCES

[1] E. ARGHIRIADE, *Remarques sur l'inverse généralize de produit de matrices*, Sci. Fis. Nat. Natur., 42 (1967), pp. 621–625.
[2] A. BEN-ISRAEL AND T.N.E. GREVILLE, *Generalized Inverses: Theory and Applications*, John Wiley, New York, 1974.
[3] S. DI AND W. SUN, *Trust region method for conic model to solve unconstrained optimization*, Optimization Methods and Software, 6 (1996), pp. 237–263.
[4] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.
[5] X. HE AND W. SUN, *Introduction to Generalized Inverses of Matrices*, Jiangsu Sci. & Tech. Publishing House, Nanjing, China, 1991.
[6] C. R. RAO AND S. K. MITRA, *Generalized Inverse of Matrices and Its Applications*, John Wiley, New York, 1971.
[7] R. J. B. SAMPAIO, W. SUN, AND J. YUAN, *On trust region algorithm for nonsmooth optimization*, Applied Math. Comput., 85 (1997), pp. 109–116.
[8] W. SUN, *Cramer rules for weighted systems*, Nanjing Daxue Xuebao Shuxue Bannian Kan, 3 (1986), pp. 117–121.

[9] W. Sun, J. Yuan, and Y. Yuan, *Trust region method of conic model for linearly constrained optimization*, J. Optim. Theory Appl., submitted.
[10] W. Sun and Y. Yuan, *Trust Region Method of Conic Model for Nonlinearly Constrained Optimization*, Math. Comp., submitted.
[11] W. Sun and Y. Yuan, *Optimization Theory and Methods*, Science Press, Beijing, 1996.
[12] G. Wang, *Perturbation theory for weighted generalized inverse*, Appl. Math. Comput. Math., 1 (1987), pp. 48–60.
[13] H. J. Werner, *G-inverse of matrix products*, in Data Analysis and Statistical Inference, S. Schach and G. Trenkler, eds., Eul Verlag, Bergisch-Gladbach, 1992, pp. 531–546.
[14] H. J. Werner, *When is B^-A^- a generalized inverse of AB?*, Linear Algebra Appl., 210 (1994), pp. 255–263.
[15] Z. Z. Yu and Z. L. Yu, *Weighted generalized inverse and rank-deficient variance of network*, J. Wuhan Surveying and Drawing College, 1 (1985).

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 776–796, July 1998

012

RELIABLE COMPUTATION OF THE CONDITION NUMBER OF A TRIDIAGONAL MATRIX IN $O(n)$ TIME*

INDERJIT S. DHILLON†

Abstract. We present one more algorithm to compute the condition number (for inversion) of an $n \times n$ tridiagonal matrix J in $O(n)$ time. Previous $O(n)$ algorithms for this task given by Higham [*SIAM J. Sci. Statist. Comput.*, 7 (1986), pp. 150–165] are based on the tempting compact representation of the upper (lower) triangle of J^{-1} as the upper (lower) triangle of a rank-one matrix. However they suffer from severe overflow and underflow problems, especially on diagonally dominant matrices. Our new algorithm avoids these problems and is as efficient as the earlier algorithms.

Key words. tridiagonal matrix, inverse, condition number, norm, overflow, underflow

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

PII. S0895479897314747

1. Introduction.

When solving a linear system $Bx = r$ we are interested in knowing how accurate the solution is. This question is often answered by showing that the solution computed in finite precision is exact for a matrix "close" to B and then measuring how sensitive the solution is to a small perturbation. The condition number of B,

$$\kappa(B) = \|B\| \cdot \|B^{-1}\|,$$

where $\|.\|$ is a matrix norm, is one such measure. It has been conjectured that the cost of computing the condition number with guaranteed accuracy is nearly the same as solving the linear system itself [10, 9]. For a dense $n \times n$ matrix B the cost of solving $Bx = r$ is $O(n^3)$, and the extra cost of computing the condition number accurately may be unacceptable. In such cases, an estimate of the condition number may be obtained at a reduced cost [15, 18].

When the coefficient matrix J is tridiagonal, the linear system $Jx = r$ may be solved in $O(n)$ time. The matrix J^{-1} is dense in general, and computation of $\kappa(J)$ by explicitly forming it would require $O(n^2)$ time. However J is completely determined by $3n-2$ parameters, and one may suspect that its inverse can be explicitly expressed in terms of an equal number of parameters. This is indeed true and J^{-1} does admit a more compact representation, namely that the upper (lower) triangle of J^{-1} is also the upper (lower) triangle of a rank-one matrix, which in turn is simply represented by the outer product of two vectors (see [3, 4, 6, 17, 21] and Theorem 2.1 below). This property of the inverse may be exploited to compute $\|J^{-1}\|_1$ and hence $\kappa_1(J)$, in $O(n)$ time; see the beginning of section 3 for details. Note that the 1-norm of a

*Received by the editors January 2, 1997; accepted for publication (in revised form) by N. J. Higham September 19, 1997; published electronically April 2, 1998. This research was supported in part, while the author was at the University of California, Berkeley, CA, by DARPA contract DAAL03-91-C-0047 through a subcontract with the University of Tennessee, DOE contract DOE-W-31-109-Eng-38 through a subcontract with Argonne National Laboratory, DOE grant DE-FG03-94ER25219, NSF grants ASC-9313958 and CDA-9401156, and DOE contract DE-AC06-76RLO 1830 through the Environmental Molecular Sciences construction project at Pacific Northwest National Laboratory (PNNL). The information presented here does not necessarily reflect the position or the policy of the U.S. Government and no official endorsement should be inferred.

http://www.siam.org/journals/simax/19-3/31474.html

†IBM Almaden Research Center, San Jose, CA 95120-6099 (dhillon@almaden.ibm.com).

matrix $B = (\beta_{ij})$ is given by

$$\|B\|_1 = \max_j \sum_i |\beta_{ij}|$$

and that $\|B\|_\infty = \|B^T\|_1$.

In [17], Higham gives three algorithms to compute $\|J^{-1}\|_\infty$ in $O(n)$ time for a general tridiagonal matrix J. However all these algorithms suffer from severe over ow and under ow problems, especially on diagonally dominant matrices. The reason for these seemingly unavoidable problems is that the intermediate quantities computed by these algorithms can vary widely in scale [20]. In this paper, we give a new algorithm that does not suffer from the above mentioned over/under ow problems. The new algorithm avoids such problems by computing sums of magnitudes of elements of the inverse itself.

For positive definite J, Higham gives another algorithm in [17] that does not suffer from over/under ow problems and is shown to be backward stable. However this algorithm is entirely different from the algorithms for a general tridiagonal. Our new algorithm works for any tridiagonal and includes positive definite J as a special case.

The paper is organized as follows. In section 2, we review the structure of the inverse of a tridiagonal matrix that enables computation of its norm in $O(n)$ time. In section 3, we present an outline of the algorithms given in [17] and show why they are unsuitable for general purpose use. We present the basic structure of our new algorithm in section 4. This algorithm works under the assumption that all principal leading and trailing submatrices are nonsingular. Section 5 sheds more light on the structure of the inverse when this assumption fails to hold. This leads to the improved algorithm of section 6, and in section 7 we give a roundoff error analysis that suggests its accuracy. This algorithm can over ow and under ow in rare cases, which is corrected by the algorithms of section 8. Accuracy of our new algorithms is confirmed by numerical results in section 10. Section 9 is a slight digression and presents an application of these algorithms for computing eigenvectors.

2. The inverse of a tridiagonal matrix. The results of this section are quite well known and are repeated here as we will frequently invoke them in later sections. A square matrix $B = (\beta_{ik})$ is called a *lower(upper) Hessenberg* matrix if $\beta_{ik} = 0$ for all pairs (i, k) such that $i + 1 < k$ $(k + 1 < i)$. Thus a lower Hessenberg matrix is nearly a lower triangular matrix but with a nonzero superdiagonal. The following theorem states that the upper half of the inverse of such a matrix admits a compact representation.

THEOREM 2.1. *Let $B = (\beta_{ik})$ be a nonsingular lower Hessenberg matrix of order n, and let $\beta_{i,i+1} \neq 0$, $i = 1, \ldots, n-1$. Then two column vectors x and y exist such that the upper half of B^{-1} equals the upper half of xy^T, i.e., $(B^{-1})_{ik} = x_i y_k$ for $i \leq k$.*

Proof. See [21]. □

Let

$$J = \begin{bmatrix} a_1 & c_1 & & & & 0 \\ b_1 & a_2 & c_2 & & & \\ & b_2 & a_3 & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & c_{n-1} \\ 0 & & & & b_{n-1} & a_n \end{bmatrix}. \tag{2.1}$$

The tridiagonal matrix given above is said to be *unreduced* or *irreducible* if $b_i \neq 0$ and $c_i \neq 0$ for all $i = 1, \ldots, n-1$. Since a tridiagonal matrix is both a lower and an upper Hessenberg matrix, we obtain the following theorem on the structure of the inverse of a tridiagonal matrix.

THEOREM 2.2. *Let J be a nonsingular unreduced tridiagonal matrix of order n. Then there exist vectors x, y, p, and q such that*

$$(J^{-1})_{ik} = \begin{cases} x_i y_k, & i \leq k, \\ p_i q_k, & i \geq k. \end{cases}$$

The vectors x and y (similarly p and q) are unique up to scaling by a nonzero factor. Note that $x_1 \neq 0$ and $y_n \neq 0$ since otherwise the entire first row or last column of J^{-1} would respectively be zero, contradicting our assumption that J is nonsingular. The above theorem seems to state that J^{-1} is determined by $4n-2$ parameters, but note that there is some redundancy in the representation of the diagonal elements since $x_i y_i = p_i q_i$ for $1 \leq i \leq n$. The following theorem makes it explicit that $3n-2$ parameters are sufficient to determine J^{-1} uniquely.

THEOREM 2.3. *Let J be a nonsingular unreduced tridiagonal matrix of order n. Then there exist vectors x and y such that*

$$(J^{-1})_{ik} = \begin{cases} x_i y_k d_k, & i \leq k, \\ y_i x_k d_k, & i \geq k, \end{cases}$$

where

$$d_1 = 1 \quad \text{and} \quad d_k = \prod_{j=1}^{k-1} \frac{c_j}{b_j}, \quad 2 \leq k \leq n.$$

Proof. The key observation is that the nonsymmetric matrix J may be written as $J = DT$, where $D = \operatorname{diag}(d_i)$ is as given above and T is symmetric. The result is then obtained by applying Theorem 2.2 to T^{-1}. See [17] for more details. □

When an off-diagonal entry is zero, it is easy to see that the "corresponding" block of the inverse is zero. For example, if $b_i = 0$ so that

$$J = \begin{bmatrix} J_1 & C_1 \\ 0 & J_2 \end{bmatrix},$$

then

$$J^{-1} = \begin{bmatrix} J_1^{-1} & X \\ 0 & J_2^{-1} \end{bmatrix},$$

where X is a rank-one matrix if $c_i \neq 0$ and zero otherwise. Note that the structure of X is consistent with Theorem 2.1.

3. Unreliability of earlier algorithms. In this section, we reproduce the three algorithms given in [17] and explain why they are unsatisfactory when implemented in finite precision. For more details on the algorithms see [16, 17].

From Theorem 2.2, the ith row sum of J^{-1} is

$$|p_i q_1| + |p_i q_2| + \cdots + |p_i q_{i-1}| + |x_i y_i| + |x_i y_{i+1}| + \cdots + |x_i y_n|,$$

Algorithm HIGHAM_1
(1) $x_1 = 1$; $x_2 = -a_1/c_1$;
  for $i = 3$ to n
    $x_i = -(a_{i-1} * x_{i-1} + b_{i-2} * x_{i-2})/c_{i-1}$;
  $y_n = 1/(b_{n-1} * x_{n-1} + a_n * x_n)$;
  $y_{n-1} = -a_n * y_n/c_{n-1}$;
  for $i = n-2$ to 1 step -1
    $y_i = -(a_{i+1} * y_{i+1} + b_{i+1} * y_{i+2})/c_i$;
(2) Repeat step (1) with x_i, y_i, b_i, and c_i replaced by q_i, p_i, c_i, and b_i, respectively.
(3) $s_n = |y_n|$;
  for $i = n-1$ to 1 step -1
    $s_i = s_{i+1} + |y_i|$;
  $t_1 = 1$;
  for $i = 2$ to n
    $t_i = t_{i-1} + |q_i|$;
  $\gamma = \max(s_1, |p_n| * t_n)$;
  for $i = 2$ to $n-1$
    $\gamma = \max(\gamma, |p_i| * t_{i-1} + |x_i| * s_i)$;
  $\|J^{-1}\|_\infty = \gamma$;

Algorithm HIGHAM_2
(1) $x_1 = 1$; $x_2 = -a_1/c_1$;
  for $i = 3$ to n
    $x_i = -(a_{i-1} * x_{i-1} + b_{i-2} * x_{i-2})/c_{i-1}$;
(2) $z_n = 1$; $z_{n-1} = -a_n/b_{n-1}$;
  for $i = n-2$ to 1 step -1
    $z_i = -(a_{i+1} * z_{i+1} + c_{i+1} * z_{i+2})/b_i$;
  $\theta = a_1 * z_1 + c_1 * z_2$;
(3) $s_n = |z_n|$;
  for $i = n-1$ to 1 step -1
    $s_i = s_{i+1} + |z_i|$;
  $t_1 = 1$;
  for $i = 2$ to $n-1$
    $t_i = t_{i-1} + |x_i|$;
  $d_1 = 1$; $\gamma = s_1$;
  for $i = 2$ to n
    $d_i = d_{i-1} * c_{i-1}/b_{i-1}$;
    $\gamma = \max(\gamma, (|z_i| * t_{i-1} + |x_i| * s_i) * |d_i|)$;
  $\gamma = \gamma/|\theta|$;
  $\|J^{-1}\|_1 = \gamma$;

FIG. 1. *Algorithms* HIGHAM_1 *and* HIGHAM_2 *compute* $\|J^{-1}\|_\infty$ *and* $\|J^{-1}\|_1$*, respectively.*

which can be simplified to

$$|p_i|(|q_1| + |q_2| + \cdots + |q_{i-1}|) + |x_i|(|y_i| + |y_{i+1}| + \cdots + |y_n|). \tag{3.1}$$

By forming the running sums

$$t_i = |q_1| + |q_2| + \cdots + |q_i|, \qquad s_i = |y_i| + |y_{i+1}| + \cdots + |y_n|,$$

all the row sums of J^{-1} may be computed in $O(n)$ time given the vectors x, y, p, and q. The vectors x and y (similarly p and q) may be computed by equating the last columns of $JJ^{-1} = I$ and the first rows of $J^{-1}J = I$ after setting x_1 to 1.

Algorithm HIGHAM_1 (see Figure 1) sets x_1 to 1 and solves $Jx = y_n^{-1}e_n$ for x and y_n. The last $n-1$ equations of $J^Ty = x_1^{-1}e_1$ are then used to solve for $y_1, \ldots, y_{n-1}$. $\|J^{-1}\|_\infty$ is then found by forming the running sums s_i, t_i and computing all the row sums using (3.1).

Algorithm HIGHAM_2 (see Figure 1) exploits Theorem 2.3 to compute $\|J^{-1}\|_1$. The vector x is computed as in the previous algorithm, and the last $n-1$ equations of $Jz = \theta e_1$ are then used to solve for $z = \theta y$. Finally the 1-norm of each column of J^{-1}, scaled by θ, is computed.

Algorithm HIGHAM_3 (see Figure 2) makes use of the LU factorization of J to solve for the first row and column of J^{-1}, which give the vectors y and p, respectively (x_1 and q_1 are set to 1). Similarly the last row and column of J^{-1} are also computed and then scaled by p_n^{-1} and y_n^{-1} to get the vectors q and x, respectively. These four vectors are then used to compute $\|J^{-1}\|_\infty$ as in Algorithm HIGHAM_1.

All of the above algorithms attempt to compute elements of the vectors x and y at some point. We show that these vectors are badly scaled especially when the matrix is diagonally dominant and, hence, well conditioned. Consider the $n \times n$ tridiagonal matrix with all diagonal elements equal to 4 and all off-diagonals equal to 1. The determinant of this matrix is asymptotical to θ^n with increasing n, where $\theta = 2 + \sqrt{3}$. By the Cauchy–Binet theorem that gives formulae for the elements of the inverse (see (4.6) below), $x_1y_1 = x_ny_n \approx \theta^{-1}$ while $|x_1y_n| \approx \theta^{-n}$. If we choose $x_1 = 1$, then $|y_n| \approx \theta^{-n}$ and $|x_n| \approx \theta^{n-1}$. The over ow threshold in double precision IEEE arithmetic is $2^{1023} \approx 10^{308}$ [2]. When $n = 540$, $\theta^{n-1} > 10^{308}$ and due

Algorithm HIGHAM_3
(1) Compute the LU factorization of J;
(2) Use the LU factorization to solve for the vectors y and z,
where $J^T y = e_1$ and $Jz = e_n$.
Similarly, solve for p and r, where $Jp = e_1$ and $J^T r = e_n$.
(3) Execute step (3) of Algorithm HIGHAM_1 with $q = p_n^{-1} r$ and $x = y_n^{-1} z$.

FIG. 2. *Algorithm* HIGHAM_3 *computes* $\|J^{-1}\|_\infty$.

to over ow all the above algorithms fail in double precision arithmetic. Note that since $|x_n/x_1| \approx \theta^{n-1}$ and $|y_n/y_1| \approx \theta^{-n+1}$, there is no choice of x_1 that can prevent over ow and under ow for all n. For the strongly diagonally dominant tridiagonal with $a_i = 1000$, $b_i = c_i = 1$, all three algorithms outlined above fail when n is only 105.

These over/under ow problems were recognized by Higham [17], [20, section 14.5], and consequently the existing LAPACK version 2.0 [1] has software only to estimate the condition number of a general tridiagonal matrix using Hager's condition estimator [15, 19]. For positive definite tridiagonals, LAPACK does contain software to accurately compute the condition number. This is based on an alternate algorithm given by Higham in [17] that is special to the positive definite case.

4. The new algorithm. As we illustrated above, the vectors x, y, p, and q that determine the inverse of a diagonally dominant matrix can be badly scaled. In this section, we present a new algorithm to compute $\|J^{-1}\|_1$ that computes sums of magnitudes of elements of J^{-1} without explicitly forming these vectors. Consequently our new algorithm does not suffer from over/under ow problems that are inevitable when x, y, p, and q are used.

Before giving all the details of our new algorithm, we illustrate the ideas on a 5×5 case. The structure of the inverse is

$$J^{-1} = \begin{bmatrix} \Delta_1 & x_1y_2 & x_1y_3 & x_1y_4 & x_1y_5 \\ p_2q_1 & \Delta_2 & x_2y_3 & x_2y_4 & x_2y_5 \\ p_3q_1 & p_3q_2 & \Delta_3 & x_3y_4 & x_3y_5 \\ p_4q_1 & p_4q_2 & p_4q_3 & \Delta_4 & x_4y_5 \\ p_5q_1 & p_5q_2 & p_5q_3 & p_5q_4 & \Delta_5 \end{bmatrix}, \quad \Delta_i \equiv x_iy_i = p_iq_i.$$

Let $s_u(i)$ denote the 1-norm of column i of the strict upper triangle of J^{-1}. Clearly

$$\begin{aligned} s_u(5) &= \left(\sum_{i=1}^{4} |x_i| \right) |y_5| \\ &= (s_u(4) + |\Delta_4|) \, \frac{|y_5|}{|y_4|}, \end{aligned}$$

and so there is a simple recurrence to build up $s_u(i)$ if Δ_i is known. Note that in the above we assumed that $y_4 \neq 0$, and, for now, we will assume that all x_i, y_i are nonzero. We can also build the following recurrence for Δ_i:

$$\Delta_{i+1} = x_{i+1}y_{i+1} \; = \; \Delta_i \, \frac{x_{i+1}}{x_i} \frac{y_{i+1}}{y_i}. \tag{4.1}$$

```
D+(1) = a1;
for i = 1 to n    1
    L+(i) = bi/D+(i); U+(i) = ci/D+(i);
    D+(i + 1) = ai+1    ci ∗ L+(i);
```

```
D−(n) = an;
for i = n    1 to 1 step    1
    U−(i) = ci/D−(i + 1); L−(i) = bi/D−(i + 1);
    D−(i) = ai    bi ∗ U−(i);
```

FIG. 3. *Algorithms to compute the triangular decompositions of J.*

Having found Δ_i, $s_u(i+1)$ may be expressed as

$$s_u(i+1) = (s_u(i) + |\Delta_i|)\,\frac{|y_{i+1}|}{|y_i|}. \tag{4.2}$$

We will see later that the ratios x_{i+1}/x_i and y_{i+1}/y_i are easily evaluated. Similarly,

$$s_l(i-1) = (s_l(i) + |\Delta_i|)\,\frac{|q_{i-1}|}{|q_i|}, \tag{4.3}$$

where $s_l(i)$ denotes the 1-norm of the ith column of the strict lower triangle of J^{-1}.

It turns out that it is possible to express the above recurrences in terms of triangular factorizations of J; two of them, as it happens. For the moment assume that the following factorizations exist:

$$J = L_+ D_+ U_+, \tag{4.4}$$

$$J = U_- D_- L_-, \tag{4.5}$$

where L_+, L_- are unit lower bidiagonal, U_+ and U_- are unit upper bidiagonal, while D_+ and D_- are diagonal matrices. Note that in the above, we use "+" to indicate a process that takes rows in increasing order while "−" indicates a process that takes rows in decreasing order. Figure 3 details the algorithms for computing these factorizations. We denote the $(i+1, i)$ element of L_+ by $L_+(i)$ and the $(i, i+1)$ element of U_- by $U_-(i)$.

In our upcoming treatment we will extensively use the famous Cauchy–Binet formula

$$B \cdot \mathrm{adj}(B) = \det(B) \cdot I, \tag{4.6}$$

where $\mathrm{adj}(B)$ is the classical *adjugate* of B and is the transpose of the matrix of cofactors [24, p. 402], to get expressions for elements of B^{-1}.

Since J is tridiagonal, (4.6) implies that

$$\Delta_i \;=\; x_i y_i \;=\; \frac{\det(J^{1:i-1}) \cdot \det(J^{i+1:n})}{\det(J)},$$

where $J^{r:s}$ denotes the principal submatrix of J in rows and columns r through s. Hence the assumption that all x_i, y_i be nonzero is identical to the assumption that the triangular factorizations (4.4) and (4.5) exist. We will remove this assumption later.

Since $L_+ e_n = e_n$ and $e_1^T L_- = e_1^T$, the first row and last column of the inverse may be expressed as

$$w_1^T \;\equiv\; e_1^T J^{-1} = e_1^T L_-^{-1} D_-^{-1} U_-^{-1} \;=\; \frac{1}{D_-(1)} e_1^T U_-^{-1},$$

$$v_n \;\equiv\; J^{-1} e_n = U_+^{-1} D_+^{-1} L_+^{-1} e_n \;=\; \frac{1}{D_+(n)} U_+^{-1} e_n.$$

Algorithm NRMINV
Compute $J = L_+D_+U_+$ and $J = U_-D_-L_-$ (see Figure 3).
$\Delta_1 = 1/D_-(1)$;
for $i = 1$ to $n-1$
    $\Delta_{i+1} = \Delta_i * \frac{D_+(i)}{D\ (i+1)}$;
$s_u(1) = 0$;
for $i = 1$ to $n-1$
    $s_u(i+1) = (s_u(i) + |\Delta_i|) * |U_-(i)|$;
$s_l(n) = 0$;
for $i = n$ to 2 step -1
    $s_l(i-1) = (s_l(i) + |\Delta_i|) * |L_+(i-1)|$;
$\gamma = 0$;
for $i = 1$ to n
    $\gamma = \max(\gamma, s_u(i) + s_l(i) + |\Delta_i|)$;
$\|J^{-1}\|_1 = \gamma$;

FIG. 4. *Algorithm* NRMINV *computes* $\|J^{-1}\|_1$.

The crucial observation is that the ratios of successive entries in w_1 and v_n are given by entries in the triangular factorizations. More precisely, the above equations may be written as

$$U_-^T w_1 = \frac{1}{D_-(1)} e_1, \tag{4.7}$$

$$U_+ v_n = \frac{1}{D_+(n)} e_n. \tag{4.8}$$

By examining the $(i+1)$st equation of (4.7) and the ith equation of (4.8), $1 \le i \le n-1$, we get

$$-U_-(i) = \frac{w_1(i+1)}{w_1(i)} = \frac{x_1 y_{i+1}}{x_1 y_i}, \tag{4.9}$$

$$-U_+(i) = \frac{v_n(i)}{v_n(i+1)} = \frac{y_n x_i}{y_n x_{i+1}}. \tag{4.10}$$

Equations (4.9) and (4.10) may now be substituted in (4.1) and (4.2) to get

$$\Delta_{i+1} = \Delta_i \frac{U_-(i)}{U_+(i)} = \Delta_i \frac{D_+(i)}{D_-(i+1)}, \quad \Delta_1 = \frac{1}{D_-(1)}, \tag{4.11}$$

$$s_u(i+1) = (s_u(i) + |\Delta_i|) \cdot |U_-(i)|, \quad s_u(1) = 0. \tag{4.12}$$

Note that the first equation of (4.7) gives $w_1(1) = \Delta_1 = 1/D_-(1)$ while the last equation of (4.8) implies that $v_n(n) = \Delta_n = 1/D_+(n)$. Similarly, we get

$$s_l(i-1) = (s_l(i) + |\Delta_i|) \cdot |L_+(i-1)|, \quad s_l(n) = 0. \tag{4.13}$$

Equations (4.11), (4.12), and (4.13) lead to Algorithm NRMINV outlined in Figure 4. This new algorithm, when implemented in finite precision, delivers correct answers on the examples of the previous section. It is also more efficient than the

TABLE 1
Comparison of arithmetic operations.

Operations $\times$ n	Divisions	Multiplications	Additions
Algorithm HIGHAM_1	4	10	7
Algorithm HIGHAM_2	3	8	5
Algorithm HIGHAM_3	7	16	16
Algorithm NRMINV	**3**	**5**	**6**

algorithms of [17]. In Table 1, we list the approximate operation counts in Algorithm NRMINV and compare them to Higham's algorithms. Note that neither U_+ nor L_- is used in Algorithm NRMINV and hence the corresponding division operations to compute them (see Figure 3) are not counted in Table 1. For more details on the operation counts for Higham's algorithms, the reader is referred to discussions of Algorithms 2, 3, and 5 in his M.Sc. thesis [16].

Recall that for our new algorithm we assumed that the factorizations in (4.4) and (4.5) exist. In the next section, we shed more light on the structure of the inverse when triangular factorization breaks down, and in section 6, we present an algorithm that handles such a breakdown.

Formula (4.11) to compute the diagonal elements of the inverse is not new and has been known for some time to researchers, especially in boundary value problems. See Meurant's survey article [22] for such formulae and more on the behavior of the inverse of a tridiagonal matrix. More recently, the diagonal of the inverse has been used to compute eigenvectors of a symmetric tridiagonal matrix [12, 23, 13, 14]. Section 9 brie y explains the connection to eigenvectors.

5. More properties of the inverse. Consider the tridiagonal matrix J of even order with $a_i = 0$ and $b_i = c_i = 1$ for all i. The factorizations (4.4) and (4.5) do not exist and all the diagonal entries of its inverse equal zero; i.e., $x_i y_i = 0$. We now present a theory that enables us to handle such a case.

THEOREM 5.1. *Let J be a nonsingular tridiagonal matrix of order n. Then $\Delta_i \equiv (J^{-1})_{ii} = 0$ if and only if either $J^{1:i-1}$ or $J^{i+1:n}$ is singular.*

Proof. This follows from (4.6) which, due to J's tridiagonal structure, implies that

$$\Delta_i \equiv (J^{-1})_{ii} = \frac{\det(J^{1:i-1}) \cdot \det(J^{i+1:n})}{\det(J)}. \qquad \square \tag{5.1}$$

Since $\Delta_i = x_i y_i$, either $x_i = 0$ or $y_i = 0$ when $\Delta_i = 0$ (note that x_i and y_i cannot both be zero if J is nonsingular because otherwise by Theorem 2.3 J^{-1} would have a zero row and column). The following theorem states that $y_i = 0$ when $J^{i+1:n}$ is singular while $x_i = 0$ when the leading submatrix $J^{1:i-1}$ is singular.

THEOREM 5.2. *Let J be a nonsingular unreduced tridiagonal matrix of order n. Then*

$$s_u(i) \equiv \sum_{k=1}^{i-1} |(J^{-1})_{k,i}| = 0 \quad \textit{if and only if} \quad J^{i+1:n} \quad \textit{is singular.} \tag{5.2}$$

Similarly,

$$s_l(i) \equiv \sum_{k=i+1}^{n} |(J^{-1})_{k,i}| = 0 \quad \textit{if and only if} \quad J^{1:i-1} \quad \textit{is singular.}$$

Proof. By the Cauchy–Binet formula in (4.6), for $k < i$,

$$\det(J) \cdot (J^{-1})_{k,i} = (-1)^{k+i}(c_k c_{k+1} \cdots c_{i-1}) \det(J^{1:k-1}) \det(J^{i+1:n}). \tag{5.3}$$

Letting $k = 1$ (take $\det(J^{1:0}) = 1$), we see that for an unreduced J, $(J^{-1})_{1,i} = 0$ if and only if $J^{i+1:n}$ is singular. The result (5.2) now follows from (5.3). ∎

Note that if $J^{1:i-1}$ and $J^{i+1:n}$ are both singular, the above theorems imply that Δ_i, $s_u(i)$ and $s_l(i)$ are zero, i.e., J^{-1} has a zero column! This leads to the following corollary.

COROLLARY 5.3. *Let J be a tridiagonal matrix of order n. If J is nonsingular, then $J^{1:i-1}$ and $J^{i+1:n}$ cannot both be singular for any $i = 2, 3, \ldots, n-1$.*

The tridiagonal structure of J is essential to the above result. To emphasize this, consider

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

where $A^{1:1}$ and $A^{3:3}$ are singular but A is not.

Now we show that for a nonsingular J no two consecutive entries in x or y can be zero. In particular this implies that both $s_u(i)$ and $s_u(i+1)$ cannot be zero.

THEOREM 5.4. *Let J be a nonsingular unreduced tridiagonal matrix of order n. Then the last (first) column or row of J^{-1} cannot have two consecutive zero entries.*

Proof. Suppose that $v_{i-1} = v_i = 0$, where $Jv = e_n$. Then the ith equation $b_{i-1}v_{i-1} + a_i v_i + c_i v_{i+1} = 0$, where $i < n$ implies that $v_{i+1} = 0$. The $(i+1)$st equation further implies that $v_{i+2} = 0$ and so on. Thus $v_{n-1} = v_n = 0$ but then the last equation $b_{n-1}v_{n-1} + a_n v_n = 1$ cannot be satisfied. ∎

The following lemma is similarly proved using the three-term recurrence for tridiagonal matrices.

LEMMA 5.5. *Let J be an unreduced (or nonsingular) tridiagonal matrix of order n. Then no two consecutive leading (or trailing) principal submatrices of J are singular.*

Proof. Suppose that $J^{1:i-1}$ and $J^{1:i}$ are singular. Then, since

$$\det(J^{1:i+1}) = a_{i+1} \det(J^{1:i}) - b_i c_i \det(J^{1:i-1})$$

and

$$-\det(J^{1:i}) + a_i \det(J^{1:i-1}) = b_{i-1}c_{i-1} \det(J^{1:i-2}),$$

$J^{1:k}$ is singular for all $k = 1, 2, \ldots, n$. But if $J^{1:1}$ is zero, then $\det(J^{1:2}) = -b_1 c_1 \neq 0$ which leads to a contradiction. ∎

We make extensive use of the following theorem in the next section.

THEOREM 5.6. *Let J be an unreduced (or nonsingular) tridiagonal matrix of order n.*

$$\textit{If } J^{1:i} \textit{ is singular, then } \det(J) = \det(J^{1:i+1}) \det(J^{i+2:n}). \tag{5.4}$$

Similarly,

$$\textit{if } J^{i:n} \textit{ is singular, then } \det(J) = \det(J^{1:i-2}) \det(J^{i-1:n}). \tag{5.5}$$

Proof. Suppose that $J^{1:i}$ is singular. Then by Lemma 5.5, $J^{1:i+1}$ is nonsingular. The Schur complement of $J^{1:i+1}$ in J is

$$\mathcal{S}(J^{1:i+1}) \;=\; J^{i+2:n} - b_{i+1}c_{i+1}e_1e_{i+1}^T(J^{1:i+1})^{-1}e_{i+1}e_1^T.$$

By Theorem 5.1, the $(i+1, i+1)$ entry of $(J^{1:i+1})^{-1}$ must be 0. Hence $\mathcal{S}(J^{1:i+1}) = J^{i+2:n}$ and

$$\det(J) \;=\; \det(J^{1:i+1})\det(\mathcal{S}(J^{1:i+1})) \;=\; \det(J^{1:i+1})\det(J^{i+2:n}). \qquad \square$$

Next we see how to detect the singularity of a leading or trailing principal submatrix. When such a submatrix is singular, triangular factorization is said to break down. However even in such a case, we can allow the computation in Figure 3 to proceed by including $\pm\infty$ in the arithmetic. We elaborate on this in the next section.

THEOREM 5.7. *Let J be an unreduced (or nonsingular) tridiagonal matrix of order n, and let D_+ and D_- be the diagonal matrices as computed by the algorithms of Figure* 3. *Then $J^{1:i}$ is singular if and only if $D_+(i) = 0$ while $J^{i:n}$ is singular if and only if $D_-(i) = 0$.*

Proof. This follows from Lemma 5.5 and the fact that

$$\det(J^{1:i-1}) \cdot D_+(i) \;=\; \det(J^{1:i}) \quad \text{and} \quad \det(J^{i+1:n}) \cdot D_-(i) \;=\; \det(J^{i:n}).$$

Note that due to Theorem 5.6, the above formulae hold even when triangular factorization "breaks down" before the computation of $D_+(i)$ or $D_-(i)$. $\square$

Finally we give an alternate formula for computing the diagonal elements of J^{-1}. Other formulae that are computationally better than (5.6) may be found in Corollary 4 of [23].

THEOREM 5.8. *Let J be a nonsingular tridiagonal matrix of order n that permits the factorizations in* (4.4) *and* (4.5). *Then $\Delta_i \equiv (J^{-1})_{ii}$ may be computed as*

$$\frac{1}{\Delta_i} = D_+(i) + D_-(i) - J_{ii}. \tag{5.6}$$

Proof. See Theorem 2 and Corollary 3 of [23]. $\square$

6. Eliminating the assumptions. In this section, we extend the algorithm outlined in section 4 to handle breakdown of triangular factorization. The theory developed in the previous section leads to these extensions.

Triangular factorizations are said to fail, or not exist, if a zero "pivot," $D_+(i)$ or $D_-(i)$, is encountered prematurely. However one of the attractions of an unreduced tridiagonal matrix is that the damage done by a zero pivot is localized. Indeed if $\pm\infty$ is added to the number system, triangular factorization cannot break down and the algorithms in Figure 3 always map J into unique L_+, D_+, U_+ and U_-, D_-, L_-. There is no need to spoil the inner loop with tests. It may no longer be true that $J = L_+D_+U_+$ or $J = U_-D_-L_-$, but equality does hold for all entries except for those at or adjacent to any infinite pivot. The IEEE arithmetic standard [2] allows such computation to proceed without breakdown, and thus we do not have to worry about zero pivots. Expressions with $\pm\infty$ are not expensive to handle if done by the hardware; see [11] for a discussion.

If $\Delta_i = 0$, i.e., $x_i = 0$ or $y_i = 0$, then equation (4.1) or (4.11) cannot be used to compute Δ_{i+1} even in exact arithmetic. Similarly $s_u(i+1)$ cannot be computed

by (4.2) or (4.12) if $s_u(i) = 0$. We now derive alternate formulae to compute Δ_{i+1} and $s_u(i+1)$ in such cases.

If $J^{1:i-1}$ is singular, i.e., $D_+(i-1) = 0$, then by (5.1) and (5.4),

$$\Delta_{i+1} \;=\; \frac{\det(J^{1:i})\det(J^{i+2:n})}{\det(J)} \;=\; \frac{\det(J^{1:i})\det(J^{i+2:n})}{\det(J^{1:i})\det(J^{i+1:n})} \;=\; \frac{1}{D_-(i+1)}, \tag{6.1}$$

and this gives a formula to compute Δ_{i+1} when the leading submatrix $J^{1:i-1}$ is singular.

Similarly if $J^{i+1:n}$ is singular, i.e., $D_-(i+1) = 0$, then by (5.1) and (5.5),

$$\Delta_{i+1} \;=\; \frac{\det(J^{1:i})\det(J^{i+2:n})}{\det(J^{1:i-1})\det(J^{i:n})} \;=\; \frac{-D_+(i)}{b_i c_i}. \tag{6.2}$$

If $J^{i+1:n}$ is singular, then y_i and $s_u(i)$ equal zero. In this case, since y_i and y_{i-1} cannot both be zero by Theorem 5.4, $s_u(i+1)$ may be computed from $s_u(i-1)$ as follows:

$$s_u(i+1) \;=\; (s_u(i-1) + |\Delta_{i-1}|)\,\frac{|y_{i+1}|}{|y_{i-1}|} + |(J^{-1})_{i,i+1}|.$$

We now simplify the above recurrence. Consider the ith equation of $J^T(x_1 y) = e_1$ when $y_i = 0$, $i \neq 1$,

$$c_{i-1}y_{i-1} + a_i y_i + b_i y_{i+1} = 0$$

$$\Rightarrow \quad \frac{y_{i+1}}{y_{i-1}} \;=\; \frac{-c_{i-1}}{b_i}.$$

Since we are considering the case when $J^{i+1:n}$ is singular, (5.3) and (5.5) imply that

$$(J^{-1})_{i,i+1} \;=\; \frac{-c_i\det(J^{1:i-1})\det(J^{i+2:n})}{\det(J)} \;=\; \frac{-c_i\det(J^{1:i-1})\det(J^{i+2:n})}{\det(J^{1:i-1})\det(J^{i:n})} \;=\; \frac{1}{b_i}.$$

Thus when $J^{i+1:n}$ is singular, $s_u(i+1)$ may be computed as

$$s_u(i+1) \;=\; (s_u(i-1) + |\Delta_{i-1}|)\,\frac{|c_{i-1}|}{|b_i|} + \frac{1}{|b_i|}. \tag{6.3}$$

$s_l(i-1)$ may similarly be computed as follows from $s_l(i+1)$ when $J^{1:i-1}$ is singular:

$$s_l(i-1) \;=\; (s_l(i+1) + |\Delta_{i+1}|)\,\frac{|b_i|}{|c_{i-1}|} + \frac{1}{|c_{i-1}|}. \tag{6.4}$$

Equations (6.1), (6.2), (6.3), and (6.4) give formulae for computing Δ_i, $s_u(i)$, and $s_l(i)$ when leading or trailing principal submatrices are exactly singular. By combining these formulae with Algorithm NRMINV of Figure 4, we get Algorithm NRMINV_NOASSUMP that is given in Figure 5. In exact arithmetic, this algorithm correctly computes the condition number of the matrix mentioned at the beginning of section 5 with $a_i = 0$, $b_i = 1$, and n even. In finite precision arithmetic, we might suspect that this algorithm breaks down when a pivot, $D_+(i)$ or $D_-(i)$, is tiny but not exactly zero. We address such issues in section 8. We now do a roundoff error analysis of our new algorithms assuming no over/under ow and indicate why they are accurate.

Algorithm NRMINV_NOASSUMP
Compute $J = L_+D_+U_+$ and $J = U_-D_-L_-$ (see Figure 3).
Set $D_+(0) = D_-(n+1) = 1$.
if $(D_+(n) = 0$ or $D_-(1) = 0)$ then $\|J^{-1}\|_1 = \infty$; return;
for $i = 2$ to $n-1$
    if $(D_+(i-1) = 0$ and $D_-(i+1) = 0)$ then $\|J^{-1}\|_1 = \infty$; return;
$\Delta_1 = 1/D_-(1)$;
for $i = 1$ to $n-1$
    if $(D_+(i) = 0$ or $D_-(i+2) = 0)$ then $\Delta_{i+1} = 0$;
    elseif $(D_-(i+1) = 0)$ then $\Delta_{i+1} = -D_+(i)/b_ic_i$;
    elseif $(D_+(i-1) = 0)$ then $\Delta_{i+1} = 1/D_-(i+1)$;
    else $\Delta_{i+1} = \Delta_i * \frac{D_+(i)}{D\ \ (i+1)}$;
$s_u(1) = 0$;
for $i = 1$ to $n-1$
    if $(D_-(i+2) = 0)$ then $s_u(i+1) = 0$;
    elseif $(D_-(i+1) = 0)$ then $s_u(i+1) = (s_u(i-1) + |\Delta_{i-1}|) * |\frac{c_i\ \ 1}{b_i}| + |\frac{1}{b_i}|$;
    else $s_u(i+1) = (s_u(i) + |\Delta_i|) * |U_-(i)|$;
$s_l(n) = 0$;
for $i = n$ to 2 step -1
    if $(D_+(i-2) = 0)$ then $s_l(i-1) = 0$;
    elseif $(D_+(i-1) = 0)$ then $s_l(i-1) = (s_l(i+1) + |\Delta_{i+1}|) * |\frac{b_i}{c_i\ \ 1}| + |\frac{1}{c_i\ \ 1}|$;
    else $s_l(i-1) = (s_l(i) + |\Delta_i|) * |L_+(i-1)|$;
$\gamma = 0$;
for $i = 1$ to n
    $\gamma = \max(\gamma, s_u(i) + s_l(i) + |\Delta_i|)$;
$\|J^{-1}\|_1 = \gamma$;

FIG. 5. *Algorithm* NRMINV_NOASSUMP *computes* $\|J^{-1}\|_1$.

7. Roundoff error analysis. We consider Algorithm NRMINV under the assumption that triangular factorization does not break down. Our model of arithmetic is that the floating point result of a basic arithmetic operation $\circ$ satisfies

$$fl(x \circ y) = (x \circ y)(1+\eta) = (x \circ y)/(1+\delta),$$

where η and δ depend on x, y, $\circ$, and the arithmetic unit but satisfy

$$|\eta| \le \varepsilon, \quad |\delta| \le \varepsilon$$

for a given ε, the latter depending only on the arithmetic unit. We shall choose freely the form (η or δ) that suits the analysis. We also adopt the convention of denoting the computed value of x by $\hat{x}$.

We now show that the computed triangular factorizations (4.4) and (4.5) are almost exact for a slightly perturbed matrix $J+\delta J$. In particular, we show that the pivots computed by the algorithms in Figure 3, $\hat{D}_+(i)$, are small relative perturbations of quantities $\widehat{D}_+(i)$ that are exact pivots for $J+\delta J_+$, where δJ_+ represents a small componentwise perturbation in the off-diagonal elements of J. Since $U_+(i) = c_i/D_+(i)$ and $L_+(i) = b_i/D_+(i)$, $\hat{U}_+(i)$ and $\hat{L}_+(i)$ can similarly be related to quantities $\widehat{U}_+(i)$

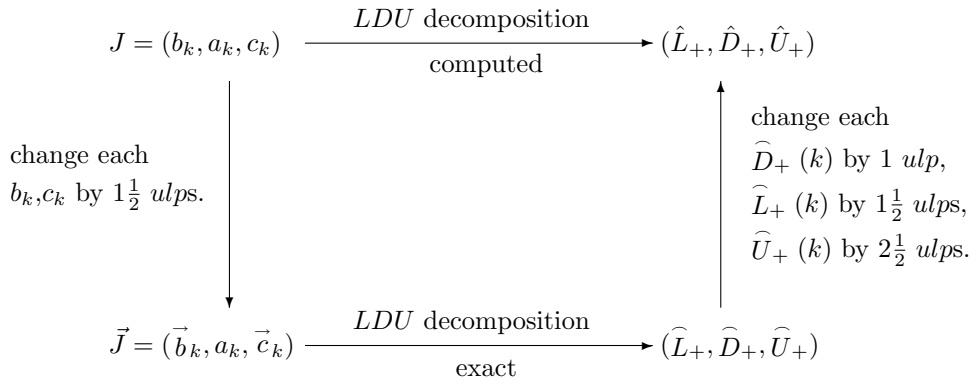


FIG. 6. *Effects of roundoff.*

and $\widehat{L}_+(i)$ that are exact for $J + \delta J_+$. An analogous result holds for the factorization $J = U_- D_- L_-$. The exact result we prove is summarized in Figure 6, where the acronym *ulp* stands for *u*nits in the *l* ast *p*lace held. It is the natural way to refer to *relative* differences between numbers. When a result is correctly rounded the error is not more than half an *ulp*.

THEOREM 7.1. *Let $J = (b_k, a_k, c_k)$ denote the tridiagonal matrix in* (2.1). *Let its LDU and UDL decompositions be computed as in Figure* 3. *In the absence of overflow and underflow, the diagram in Figure* 6 *commutes, and, for each k, $\hat{D}_+(k)$ differs from $\widehat{D}_+(k)$ by* 1 *ulp, $\hat{L}_+(k)$, $\hat{U}_+(k)$ differ from $\widehat{L}_+(k)$, $\widehat{U}_+(k)$ by $1\frac{1}{2}$ and $2\frac{1}{2}$ ulps, respectively, while $\vec{b}_k$, $\vec{c}_k$ differ from b_k, c_k by $1\frac{1}{2}$ ulps each. A similar result holds for the UDL factorization.*

Proof. We write down the exact equations satisfied by the computed quantities:

$$\hat{L}_+(k-1) = \frac{b_{k-1}}{\hat{D}_+(k-1)}(1+\varepsilon_/),$$

$$\hat{U}_+(k-1) = \frac{c_{k-1}}{\hat{D}_+(k-1)}(1+\varepsilon_{//}),$$

$$\hat{D}_+(k) = \left(a_k - c_{k-1}\hat{L}_+(k-1)\cdot(1+\varepsilon_*)\right)/(1+\varepsilon_k),$$

$$\Rightarrow \quad (1+\varepsilon_k)\hat{D}_+(k) = a_k - \frac{b_{k-1}c_{k-1}}{\hat{D}_+(k-1)}(1+\varepsilon_*)(1+\varepsilon_/). \tag{7.1}$$

In the above, all the ε depend on k but we have chosen to single out the one that accounts for the subtraction as it is the only one where the dependence on k must be made explicit. We now introduce the quantities

$$\widehat{D}_+(k) = \hat{D}_+(k)(1+\varepsilon_k), \tag{7.2}$$

$$\vec{b}_{k-1} = b_{k-1}\sqrt{(1+\varepsilon_*)(1+\varepsilon_/)(1+\varepsilon_{k-1})}, \tag{7.3}$$

$$\vec{c}_{k-1} = c_{k-1}\sqrt{(1+\varepsilon_*)(1+\varepsilon_/)(1+\varepsilon_{k-1})}. \tag{7.4}$$

Substituting (7.2), (7.3), and (7.4) in (7.1), we see that $\widehat{D}_+(k)$ is exact for $\vec{J} = [\vec{b}_k, a_k, \vec{c}_k]$, i.e.,

$$\widehat{D}_+(k) = a_k - \frac{\vec{b}_{k-1}\vec{c}_{k-1}}{\widehat{D}_+(k-1)}.$$

To satisfy the exact mathematical relations

$$\widehat{L}_+(k) = \frac{\vec{b}_k}{\widehat{D}_+(k)}, \quad \widehat{U}_+(k) = \frac{\vec{c}_k}{\widehat{D}_+(k)},$$

we set

$$\widehat{L}_+(k) = \hat{L}_+(k)\sqrt{\frac{1+\varepsilon_*}{(1+\varepsilon_k)(1+\varepsilon_/)}},$$

$$\widehat{U}_+(k) = \hat{U}_+(k)\frac{1}{1+\varepsilon_{//}}\sqrt{\frac{(1+\varepsilon_*)(1+\varepsilon_/)}{1+\varepsilon_k}},$$

and the result holds. The result for the factorization $J = U_- D_- L_-$ is similarly proved. □

The observant reader would have noted that the above is not a pure backward error analysis. We have put small perturbations not only on the input but also on the output. This property is called mixed stability in [7], but note that our perturbations are relative ones.

It is important to note that the backward perturbations for the LDU factorization differ from the ones for the UDL factorization. By (4.11), Δ_i is formed by a ratio of $D_+(i)$ and $D_-(i+1)$. Since this mixes the LDU and UDL decompositions, the roundoff error analysis given above *does not* enable us to relate the computed value of all the Δ_i to a *single* perturbed tridiagonal matrix. However if small relative changes to the off-diagonal entries of J lead to "small" changes in its LDU and UDL factorizations, then Theorem 7.1 implies that Algorithm NRMINV "accurately" computes the condition number of J. The latter implication is easily seen to be true by observing that the quantities Δ_i, $s_u(i)$, $s_l(i)$ are computed from the LDU and UDL factorizations by multiplications, divisions, and additions of nonnegative numbers. The case of Algorithm NRMINV_NOASSUMP is similar.

Often the triangular factorizations (4.4) and (4.5) can be very sensitive to small changes in the entries of the tridiagonal matrix. These are precisely the situations when a submatrix of J is close to being singular and there is element growth in the factorizations. Thus we may suspect that our algorithm delivers inaccurate answers in such cases. However numerical experience, given in section 10, indicates that the condition number is computed accurately despite element growth. It is an open problem to explain this phenomenon. We feel the situation is somewhat similar to Algorithms HIGHAM_1 and HIGHAM_2 that were outlined in section 3. In [17], Higham observes that when the latter algorithms do not over/under ow their answers are very accurate, but no error analysis has been able to explain this accuracy. One approach to proving accuracy of our algorithm may be to relate both sets of pivots, D_+ and D_-, to a *single perturbed matrix.*

8. Handling overflow and underflow. Algorithm NRMINV_NOASSUMP also suffers from the limited range of numbers that can be represented in a digital computer. Consider the matrix

$$J \;=\; \begin{bmatrix} 1000 & 100 \\ 100 & 10^{-306} \end{bmatrix}.$$

In its UDL decomposition, $D_-(2) = 10^{-306}$ while $D_-(1)$ is computed as

$$D_-(1) = 1000 - \frac{10^4}{10^{-306}} = 1000 - 10^{310}.$$

In IEEE double precision arithmetic, the above value over ows and $D_-(1)$ is set to $-\infty$ [2]. Since $\Delta_1 = 1/D_-(1)$, it is computed to be 0 by Algorithm NRMINV_NOASSUMP. Δ_2 is then computed as

$$\Delta_2 \;=\; \Delta_1 \cdot \left(\frac{D_+(1)}{D_-(2)}\right) \;=\; 0 \cdot \left(\frac{1000}{10^{-306}}\right).$$

Again the value $1000/10^{-306}$ over ows and Δ_2 is set to $0 \cdot \infty =$ Not a Number (NaN). Note that J is perfectly well conditioned with $\Delta_2 = -0.1$, and

$$J^{-1} \;\approx\; \begin{bmatrix} -10^{-310} & 0.01 \\ 0.01 & -0.1 \end{bmatrix}.$$

Thus Algorithm NRMINV_NOASSUMP malfunctions due to over ow problems. Under- ow in computing Δ_i by (4.11) can cause similar problems.

We now show how to overcome such over ow and under ow. Before doing so we emphasize that the above over/under ow problems are not as severe as those in the algorithms of [17]. The discerning reader would have noticed that problems in the earlier algorithms are inevitable due to the explicit computation of the vectors x, y, p, and q; see section 3 for more details.

There are two problems that we must address. The first is to avoid NaNs in the computation. A NaN results when evaluating expressions such as $0 \cdot \infty$, $\frac{0}{0}$, and $\frac{\infty}{\infty}$. Algorithm NRMINV_FINAL1 given in Figure 7 prevents the formation of NaNs by explicitly avoiding such expressions and handling separately the special cases when $D_+(i)$ or $D_-(i)$ equals 0 or ∞.

The second difficulty occurs if Δ_i over ows or under ows to 0 when computed as

$$\Delta_i \;=\; \Delta_{i-1} \frac{D_+(i-1)}{D_-(i)}.$$

It is incorrect to use such a Δ_i to compute Δ_{i+1} by the above recurrence. We solve this problem by computing Δ_{i+1} as

$$\Delta_{i+1} \;=\; 1/\left(D_-(i+1) - \frac{b_i c_i}{D_+(i)}\right) \tag{8.1}$$

in such a case. The above formula is a consequence of Theorem 5.8. Note that (8.1) leads to the correct value of Δ_{i+1} when $D_+(i-1) = 0$ or $D_-(i+1) = 0$; see (6.1) and (6.2).

Thus Algorithm NRMINV_FINAL1 tries to cure the over/under ow problems, and we have found its computer implementation to be accurate on all tridiagonal matrices

Algorithm NRMINV_FINAL1
Compute $J = L_+D_+U_+$ and $J = U_-D_-L_-$ as follows:
$D_+(1) = a_1$;
for $i = 1$ to $n-1$
    if ($D_+(i) = 0$ and $b_ic_i = 0$) then $\|J^{-1}\|_1 = \infty$; return;
    else $L_+(i) = b_i/D_+(i)$; $D_+(i+1) = a_{i+1} - c_i * L_+(i)$;
$D_-(n) = a_n$;
for $i = n-1$ to 1 step -1
    if ($D_-(i+1) = 0$ and $b_ic_i = 0$) then $\|J^{-1}\|_1 = \infty$; return;
    else $U_-(i) = c_i/D_-(i+1)$; $D_-(i) = a_i - b_i * U_-(i)$;
if ($D_+(n) = 0$ or $D_-(1) = 0$) then $\|J^{-1}\|_1 = \infty$; return;
for $i = 2$ to $n-1$
    if ($D_+(i-1) = 0$ and $D_-(i+1) = 0$) then $\|J^{-1}\|_1 = \infty$; return;
$\Delta_1 = 1/D_-(1)$;
for $i = 1$ to $n-1$
    if ($D_+(i) = 0$ or $1/D_-(i+1) = 0$) then $\Delta_{i+1} = 0$;
    elseif ($D_-(i+1) = 0$) then $\Delta_{i+1} = -D_+(i)/b_ic_i$;
    elseif ($1/D_+(i) = 0$) then $\Delta_{i+1} = 1/D_-(i+1)$;
    elseif ($\Delta_i = 0$) then $\Delta_{i+1} = 1/\left(D_-(i+1) - \frac{b_ic_i}{D_+(i)}\right)$;
    else $\Delta_{i+1} = \frac{\Delta_i}{D\ (i+1)} * D_+(i)$;
    if ($1/\Delta_{i+1} = 0$) then $\|J^{-1}\|_1 = \infty$; return;
$s_u(1) = 0$;
for $i = 1$ to $n-1$
    if ($s_u(i) + |\Delta_i| = 0$) then
        $s_u(i+1) = (s_u(i-1) + |\Delta_{i-1}|) * |\frac{c_i\ _1}{b_i}| + |\frac{1}{b_i}|$;
    else $s_u(i+1) = (s_u(i) + |\Delta_i|) * |U_-(i)|$;
    if ($1/s_u(i+1) = 0$) then $\|J^{-1}\|_1 = \infty$; return;
$s_l(n) = 0$;
for $i = n$ to 2 step -1
    if ($s_l(i+1) + |\Delta_{i+1}| = 0$) then
        $s_l(i-1) = (s_l(i+1) + |\Delta_{i+1}|) * |\frac{b_i}{c_i\ _1}| + |\frac{1}{c_i\ _1}|$;
    else $s_l(i-1) = (s_l(i) + |\Delta_i|) * |L_+(i-1)|$;
    if ($1/s_l(i-1) = 0$) then $\|J^{-1}\|_1 = \infty$; return;
$\gamma = 0$;
for $i = 1$ to n
    if ($s_u(i) + s_l(i) + |\Delta_i| > \gamma$) then
        $\gamma = s_u(i) + s_l(i) + |\Delta_i|$;
$\|J^{-1}\|_1 = \gamma$;

FIG. 7. *Algorithm* NRMINV_FINAL1 *computes* $\|J^{-1}\|_1$.

in our test-bed. Numerical results to show this are presented in the next section. In addition, this algorithm also works for tridiagonal matrices that are not unreduced, i.e., where some of the off-diagonal entries may be zero. None of the elaborate techniques used in [16, 17] are needed to handle this special case. As written, the algorithm requires IEEE arithmetic but it is easily modified to prevent over ow.

In spite of the above precautions, Algorithm NRMINV_FINAL1 can march danger-

ously close to the over ow and under ow thresholds. When a pivot element $D_+(i)$ or $D_-(i)$ is tiny, intermediate quantities can vary widely in magnitude while computing $s_u(i)$ and $s_l(i)$ by (4.12) and (4.13). We now present an alternate algorithm that tries to avoid large intermediate numbers. To avoid division by the tiny pivot $D_-(i+1)$ in (4.12), we may write $s_u(i+1)$ in terms of $s_u(i-1)$ as follows:

$$\begin{aligned} s_u(i+1) &= (s_u(i) + |\Delta_i|) \left| \frac{c_i}{D_-(i+1)} \right| \\ &= (s_u(i-1) + |\Delta_{i-1}|) \left| \frac{c_{i-1}c_i}{D_-(i)D_-(i+1)} \right| + \left| \frac{\Delta_i c_i}{D_-(i+1)} \right|. \end{aligned} \tag{8.2}$$

Now, the formula for computing $D_-(i)$ (see Figure 3) implies that

$$D_-(i+1)D_-(i) = D_-(i+1)(a_i - b_i c_i / D_-(i+1)) = D_-(i+1)a_i - b_i c_i, \tag{8.3}$$

and using (5.6),

$$\frac{\Delta_i c_i}{D_-(i+1)} \;=\; \frac{c_i}{D_-(i+1)D_+(i) - b_i c_i}. \tag{8.4}$$

Substitution of (8.3) and (8.4) in (8.2) leads to the desired formula

$$s_u(i+1) = (s_u(i-1) + |\Delta_{i-1}|) \left| \frac{c_{i-1}c_i}{D_-(i+1)a_i - b_i c_i} \right| + \left| \frac{c_i}{D_-(i+1)D_+(i) - b_i c_i} \right|. \tag{8.5}$$

Unlike (4.12), the above formula does not involve division by the tiny pivot element $D_-(i+1)$. Thus no large intermediate quantities are formed. Similarly, $s_l(i-1)$ may be expressed in terms of $s_l(i+1)$ to avoid division by a small $D_+(i-1)$. Note that in the extreme case when $D_-(i+1) = 0$, (8.5) simplifies to (6.3). Equation (8.5) can alternatively be obtained by taking the 2×2 matrix

$$\begin{bmatrix} a_i & c_i \\ b_i & D_-(i+1) \end{bmatrix}$$

as a pivot in block Gaussian Elimination (instead of $D_-(i+1)$) and using the corresponding block $U_- D_- L_-$ factorization to compute $s_u(i+1)$. When $D_-(i+1)$ is tiny, it can be shown that using this 2×2 pivot prevents element growth unless J is nearly singular. Algorithm NRMINV_FINAL2 given in Figure 8 uses such a pivot strategy to compute $s_u(i)$ and $s_l(i)$. Also note that in Algorithm NRMINV_FINAL2 we use (5.6) instead of (4.11) to compute Δ_i.

Although Algorithm NRMINV_FINAL2 tends to have less element growth in its computation, it is not clear whether it is more accurate than Algorithm NRMINV_FINAL1. Numerical experience, given in section 10, indicates that both these algorithms are accurate. Our personal preference is for Algorithm NRMINV_FINAL2 since the intermediate quantities computed by it do not vary widely in scale.

9. Another application. In computing $\|J^{-1}\|$, we need to find the column of J^{-1} with the largest 1-norm. We now brie y mention another application where we may need to identify such a column.

Given a real, symmetric tridiagonal matrix T and an accurate approximation to an eigenvalue $\hat{\lambda}$, we can attempt to find the corresponding eigenvector by solving

$$(T - \hat{\lambda} I) z_k = e_k,$$

where e_k is the kth column of the identity matrix (the above may also be thought

Algorithm NRMINV_FINAL2
Compute $J = L_+D_+U_+$ and $J = U_-D_-L_-$ as follows:
$D_+(1) = a_1$;
for $i = 1$ to $n-1$
    if ($D_+(i) = 0$ and $b_ic_i = 0$) then $\|J^{-1}\|_1 = \infty$; return;
    else $L_+(i) = b_i/D_+(i)$; $D_+(i+1) = a_{i+1} - c_i * L_+(i)$;
$D_-(n) = a_n$;
for $i = n-1$ to 1 step -1
    if ($D_-(i+1) = 0$ and $b_ic_i = 0$) then $\|J^{-1}\|_1 = \infty$; return;
    else $U_-(i) = c_i/D_-(i+1)$; $D_-(i) = a_i - b_i * U_-(i)$;
if ($D_+(n) = 0$ or $D_-(1) = 0$) then $\|J^{-1}\|_1 = \infty$; return;
for $i = 2$ to $n-1$
    if ($D_+(i-1) = 0$ and $D_-(i+1) = 0$) then $\|J^{-1}\|_1 = \infty$; return;
$\Delta_1 = 1/D_-(1)$;
for $i = 1$ to $n-1$
    $\Delta_{i+1} = 1/(D_-(i+1) - \frac{b_ic_i}{D_+(i)})$;
    if ($1/\Delta_{i+1} = 0$) then $\|J^{-1}\|_1 = \infty$; return;
$s_u(1) = 0$;
for $i = 1$ to $n-1$
    $\mathrm{DET} = D_-(i+1)a_i - b_ic_i$;
    if ($1/D_-(i+1) = 0$ or $|D_-(i+1) \cdot a_i| \geq |\mathrm{DET}|$) then
        $s_u(i+1) = (s_u(i) + |\Delta_i|) * |U_-(i)|$;
    else
        $s_u(i+1) = (s_u(i-1) + |\Delta_{i-1}|) * |\frac{c_i \;\; _1c_i}{\mathrm{DET}}| + |\frac{c_i}{|D \;\; _{(i+1)}D_+(i) - b_ic_i}|$;
    if ($1/s_u(i+1) = 0$) then $\|J^{-1}\|_1 = \infty$; return;
$s_l(n) = 0$;
for $i = n$ to 2 step -1
    $\mathrm{DET} = D_+(i-1)a_i - b_{i-1}c_{i-1}$;
    if ($1/D_+(i-1) = 0$ or $|D_+(i-1) \cdot a_i| \geq |\mathrm{DET}|$) then
        $s_l(i-1) = (s_l(i) + |\Delta_i|) * |L_+(i-1)|$;
    else
        $s_l(i-1) = (s_l(i+1) + |\Delta_{i+1}|) * |\frac{b_i \;\; _1b_i}{\mathrm{DET}}| + |\frac{b_i \;\; _1}{D_+(i-1)D \;\; _{(i)} - b_i \;\; _1c_i \;\; _1}|$;
    if ($1/s_l(i-1) = 0$) then $\|J^{-1}\|_1 = \infty$; return;
$\gamma = 0$;
for $i = 1$ to n
    if ($s_u(i) + s_l(i) + |\Delta_i| > \gamma$) then
        $\gamma = s_u(i) + s_l(i) + |\Delta_i|$;
$\|J^{-1}\|_1 = \gamma$;

FIG. 8. *Algorithm* NRMINV_FINAL2 *computes* $\|J^{-1}\|_1$.

of as the first step of inverse iteration with e_k as the starting vector). However an arbitrary choice of k does not always work, as observed by Wilkinson in [25, 26]. Note that the pair $(\hat{\lambda}, z_k)$ has the residual norm

$$\frac{\|(T - \hat{\lambda}I)z_k\|}{\|z_k\|} = \frac{1}{\|(T - \hat{\lambda}I)^{-1}e_k\|}, \tag{9.1}$$

TABLE 2
Test matrices.

Matrix Type	Description
1	Nonsymmetric random tridiagonal — each element is uniformly distributed in the interval $[\ \ 1, 1]$.
2	Symmetric tridiagonal $J = Q^T DQ$ with Q random orthogonal and D diagonal with one element equal to 1 and all others equal to ϵ.
3	Symmetric tridiagonal $J = Q^T DQ$ with Q random orthogonal and D diagonal with one element equal to ϵ and all others equal to 1.
4	Symmetric tridiagonal $J = Q^T DQ$ with Q random orthogonal and D diagonal with elements geometrically distributed from ϵ to 1.
5	Symmetric tridiagonal $J = Q^T DQ$ with Q random orthogonal and D diagonal with elements uniformly distributed from ϵ to 1.
6	Symmetric Toeplitz tridiagonal with $a_i = 64$, $b_i = c_i = 1$.
7	Symmetric Toeplitz tridiagonal with $a_i = 10^8$, $b_i = c_i = 1$.
8	Symmetric Toeplitz tridiagonal with $a_i = 0$, $b_i = c_i = 1$.
9	Nonsymmetric random tridiagonal as in Type 1 but with some off-diagonals set to zero.

TABLE 3
Computation of $\kappa(J) = \|J\|_1 \cdot \|J^{-1}\|_1$ *on matrices of order* 41.

Matrix Type	$\kappa(J)$ computed by				
	Algorithm NRMINV_FINAL1	Algorithm NRMINV_FINAL2	Algorithm HIGHAM1	Algorithm HIGHAM2	LAPACK's condition estimator (DGTCON)
1	111.9	111.9	111.9	111.9	109.7
2	$7.5 \cdot 10^{15}$	$7.6 \cdot 10^{15}$	NaN	NaN	$7.6 \cdot 10^{15}$
3	$5.4 \cdot 10^{15}$	$5.4 \cdot 10^{15}$	NaN	NaN	$5.4 \cdot 10^{15}$
4	$6.3 \cdot 10^{15}$	$6.3 \cdot 10^{15}$	$6.3 \cdot 10^{15}$	$6.3 \cdot 10^{15}$	$6.3 \cdot 10^{15}$
5	$7.6 \cdot 10^{15}$	$7.7 \cdot 10^{15}$	$7.7 \cdot 10^{15}$	$7.6 \cdot 10^{15}$	$7.7 \cdot 10^{15}$
6	1.06	1.06	1.06	1.06	1.06
7	1.0	1.0	NaN	NaN	1.0
8	∞	∞	NaN	∞	∞
9	$1.3 \cdot 10^3$	$1.3 \cdot 10^3$	NaN	NaN	$1.3 \cdot 10^3$

where we assume that $\hat{\lambda}$ is not an exact eigenvalue of T. The goal is to obtain a small residual norm, but an arbitrary choice of k fails because not every column of $(T - \hat{\lambda}I)^{-1}$ is large in magnitude. However when $\hat{\lambda}$ is close to an eigenvalue, there must exist a column k of $(T - \hat{\lambda}I)^{-1}$ that has a large norm. The corresponding pair $(\hat{\lambda}, z_k)$ has a small residual norm, and it can be shown that z_k is close to an eigenvector. The optimal choice of k minimizes the residual norm (9.1), i.e., it maximizes $\|(T - \hat{\lambda}I)^{-1}e_k\|$. Thus the algorithms discussed earlier in the paper provide a solution to this problem in $O(n)$ time. Algorithms NRMINV_FINAL1 and NRMINV_FINAL2 are easily modified to give the solution when the 2-norm is considered.

Often when the corresponding eigenvalue is sufficiently isolated, it suffices to choose k such that the (k, k) entry of $(T - \hat{\lambda}I)^{-1}$ has the largest absolute value among all diagonal elements of the inverse. For more on this problem, the interested reader is referred to [14, 23] and [12, Chapter 3]. As a way to find an optimal k, Jesse Barlow [5] also independently discovered recurrences similar to (4.11), (4.12), and (4.13).

10. Numerical results. In this section, we present numerical results of our new algorithms and compare them with existing algorithms. A variety of tridiagonal matrices listed in Table 2 forms our test-bed. The matrices of type 2–5 were obtained by Householder reduction of a random dense symmetric matrix that had the desired spectrum. See [8] for more on the generation of such matrices.

TABLE 4
Computation of $\kappa(J) = \|J\|_1 \cdot \|J^{-1}\|_1$ on matrices of order 200.

Matrix Type	$\kappa(J)$ computed by				
	Algorithm NRMINV_FINAL1	Algorithm NRMINV_FINAL2	Algorithm HIGHAM1	Algorithm HIGHAM2	LAPACK's condition estimator (DGTCON)
1	$1.9 \cdot 10^3$	$1.9 \cdot 10^3$	$1.9 \cdot 10^3$	$1.9 \cdot 10^3$	$1.2 \cdot 10^3$
2	$7.4 \cdot 10^{15}$	$7.5 \cdot 10^{15}$	NaN	NaN	$7.4 \cdot 10^{15}$
3	$4.5 \cdot 10^{15}$	$4.5 \cdot 10^{15}$	NaN	NaN	$4.5 \cdot 10^{15}$
4	$9.9 \cdot 10^{15}$	$9.9 \cdot 10^{15}$	$9.9 \cdot 10^{15}$	$9.9 \cdot 10^{15}$	$9.9 \cdot 10^{15}$
5	$1.2 \cdot 10^{16}$	$1.2 \cdot 10^{16}$	$1.2 \cdot 10^{16}$	$1.2 \cdot 10^{16}$	$1.2 \cdot 10^{16}$
6	1.06	1.06	1.06	1.06	1.06
7	1.0	1.0	NaN	NaN	1.0
8	200.0	200.0	200.0	200.0	2.0
9	$2.0 \cdot 10^3$	$2.0 \cdot 10^3$	NaN	NaN	$2.0 \cdot 10^3$

TABLE 5
Timing results.

Matrix Type	Time taken by LAPACK's DGTCON (in ms.)			Time(DGTCON) / Time(NRMINV_FINAL1)			Time(DGTCON) / Time(NRMINV_FINAL2)		
	$n = 41$	$n = 200$	$n = 1000$	$n = 41$	$n = 200$	$n = 1000$	$n = 41$	$n = 200$	$n = 1000$
1	0.2	1.1	5.4	2.0	1.6	1.6	1.0	1.6	1.6
2	0.3	1.1	5.4	3.0	1.6	1.7	3.0	1.6	1.6
3	0.2	1.1	5.4	2.0	1.8	1.6	1.0	1.6	1.5
4	0.2	1.1	5.5	2.0	1.8	1.7	1.0	1.6	1.6
5	0.3	1.1	5.5	3.0	1.6	1.7	1.5	1.6	1.6
6	0.3	1.6	7.2	3.0	2.3	2.2	1.5	2.0	1.8
7	0.3	1.5	6.8	3.0	2.5	2.1	3.0	1.9	1.7
8	0.0	0.9	4.6	1.0	1.5	1.5	1.0	1.5	1.5
9	0.3	1.1	5.3	1.5	1.6	1.6	3.0	2.2	1.9

The results given in Tables 3 and 4 support our claim that the algorithms in [17] are susceptible to severe over ow and under ow problems. However they produce accurate answers when they do not suffer from such problems. The new algorithms outlined in the previous section, Algorithm NRMINV_FINAL1 and Algorithm NRMINV_FINAL2, give accurate answers on all our test matrices. Both the algorithms appear to be comparable in accuracy. In our numerical results, we have also included the current algorithm in LAPACK that estimates the condition number of a tridiagonal matrix [19]. This algorithm is guaranteed to give a lower bound on the condition number, and extensive testing done in [19] indicates that its estimates are good approximations to the exact condition number in most cases. For all our test matrices, except one, the condition numbers are estimated accurately. The only exception is the Toeplitz matrix with 0 on the diagonals and 1 on the off-diagonals; see Table 4. This example is similar to the one given in [19, p. 386], and LAPACK's condition estimator underestimates its condition number by a factor of $n/2$ for $n = 200$.

In Table 5, we compare the times taken by our new algorithms with LAPACK's condition estimator. The latter also appears to take $O(n)$ time but our new algorithms are up to three times faster. These timing experiments were conducted on an IBM RS/6000 processor.

11. Conclusions. In this paper, we have given stable algorithms to compute the condition number of a tridiagonal matrix in $O(n)$ time. Algorithm NRMINV (see Figure 4) contains the main new ideas and forms the basis of Algorithms NRMINV_FINAL1 and NRMINV_FINAL2 (see Figures 7 and 8). The latter algorithms may be directly implemented to give reliable numerical software and do not suffer from the inherent over/under ow problems of the earlier algorithms presented in [17].

Acknowledgments. I would like to thank Professors B. N. Parlett, J. W. Demmel, and N. J. Higham for a careful reading of the manuscript and for many useful suggestions.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, LAPACK Users' Guide, 2nd ed., SIAM, Philadelphia, PA, 1995.

[2] ANSI/IEEE, New York, *IEEE Standard for Binary Floating Point Arithmetic*, Std 754-1985 ed., 1985.

[3] Edgar Asplund, *Inverses of matrices* $\{a_{ij}\}$ *which satisfy* $a_{ij} = 0$ *for* $j > i + p$, Math. Scand., 7 (1959), pp. 57–60.

[4] S. O. Asplund, *Finite boundary value problems solved by Green's matrix*, Math. Scand., 7 (1959), pp. 49–56.

[5] J. Barlow, *Private communication,* Pennsylvania State University, University Park, PA, 1996.

[6] B. Bukhberger and G. A. Emel'yanenko, *Methods of inverting tridiagonal matrices*, Comput. Math. Math. Phys., 13 (1973), pp. 10–20.

[7] L. S. DeJong, *Towards a formal definition of numerical stability*, Numer. Math., 28 (1977), pp. 211–220.

[8] J. Demmel and A. McKenney, *A Test Matrix Generation Suite*, LAPACK Working Note #9, Technical report, Courant Institute, Computer Science Department, New York, 1989.

[9] J. W. Demmel, *The Complexity of Condition Estimation*, manuscript, University of California, Berkeley, CA, 1997.

[10] James W. Demmel, *Open Problems in Numerical Linear Algebra*, LAPACK Working Note 47, IMA Preprint Series #961, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, MN, April 1992.

[11] James W. Demmel and Xiaoye Li, *Faster numerical algorithms via exception handling*, IEEE Trans. Comput., 43 (1994), pp. 983–992.

[12] I. S. Dhillon, *A New* $O(n^2)$ *Algorithm for the Symmetric Tridiagonal Eigenvalue/Eigenvector Problem*, Ph.D. thesis, Computer Science Division, EECS Dept., University of California, Berkeley, CA, May 1997; also available as Computer Science Division Technical report UCB//CSD-97-971.

[13] I. S. Dhillon and B. N. Parlett, *Orthogonal Eigenvectors Without Gram-Schmidt*, 1997, manuscript.

[14] K. V. Fernando, *On computing an eigenvector of a tridiagonal matrix. Part* I: *Basic results*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 1013-1034.

[15] W. W. Hager, *Condition estimators*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 311–316.

[16] N. J. Higham, *Matrix condition numbers*, M.Sc. thesis, Dept. of Mathematics, University of Manchester, Manchester, England, 1983.

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

[18] N. J. Higham, *A survey of condition number estimation for triangular matrices*, SIAM Rev., 29 (1987), pp. 575–596.

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

[20] N. J. Higham, *Accuracy and stability of numerical algorithms*, SIAM, Philadelphia, PA, 1996.

[21] I. Ikebe, *On inverses of Hessenberg matrices*, Linear Algebra Appl., 24 (1979), pp. 93–97.

[22] G. Meurant, *A review on the inverse of symmetric tridiagonal and block tridiagonal matrices*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 707–728.

[23] B. N. Parlett and I. S. Dhillon, *Fernando's solution to Wilkinson's problem: an application of double factorization*, Linear Algebra Appl., 267 (1997), pp. 247–279.

[24] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[25] J. H. Wilkinson, *The calculation of the eigenvectors of codiagonal matrices*, Computer J., 1 (1958), pp. 90–96.

[26] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 797–806, July 1998

013

A FAST ALGORITHM FOR INVERSION OF CONFLUENT VANDERMONDE-LIKE MATRICES INVOLVING POLYNOMIALS THAT SATISFY A THREE-TERM RECURRENCE RELATION*

XU ZHONG† AND YOU ZHAOYONG‡

Abstract. The present paper describes a new fast algorithm for inversion of confluent Vandermonde-like matrices. Our algorithm generalizes a scheme presented by Calvetti and Reichel [27] for fast inversion of Vandermonde-like matrices.

Key words. confluent Vandermonde-like matrix, inverse, fast algorithm

AMS subject classifications. 65F05, 15A09

PII. S089547989630749X

1. Introduction.

Let $\alpha_1, \alpha_2, \ldots, \alpha_p$ be a set of distinct nodes, let $n_1, n_2, \ldots, n_p$ be p positive integers, and let $n = \sum_{i=1}^{p} n_i$. Let $\{p_k(x)\}_{k=0}^{n}$ be a family of polynomials that satisfy a three-term recurrence relation

$$\begin{aligned} p_0(x) &= 1, \qquad p_1(x) = \theta_1(x - \beta_1)p_0(x), \\ p_k(x) &= \theta_k(x - \beta_k)p_{k-1}(x) - \gamma_k p_{k-2}(x), \qquad (k \geq 2), \end{aligned} \tag{1.1}$$

where $\theta_k \neq 0$ for all k. The n × n con uent Vandermonde-like matrix is given by

$$V_c = (\mathbf{v}(\alpha_1), \mathbf{v}'(\alpha_1), \ldots, \mathbf{v}^{(n_1-1)}(\alpha_1), \ldots, \mathbf{v}(\alpha_p), \mathbf{v}'(\alpha_p), \ldots, \mathbf{v}^{(n_p-1)}(\alpha_p)), \tag{1.2}$$

where $\mathbf{v}(x) = (p_0(x), p_1(x), \ldots, p_{n-1}(x))^T$. If $p_k(x) = x^k$, then the matrix V_c reduces to a con uent Vandermonde matrix [2, 9, 28]. When the positive integers n_j satisfy $n_1 = n_2 = \cdots = n_p = 1$, the matrix V_c is the Vandermonde-like matrix [20, 27]. In the case of $n_1 = n_2 = \cdots = n_p = 1$ and $p_k(x) = x^k$, the matrix V_c yields the well-known Vandermonde matrix. The associated con uent Vandermonde-like systems $V_c a = f$ and $V_c^T a = f$, where f is a given right-hand side vector, arise in a variety of applications such as polynomial interpolation and approximation of linear functionals [5, 6].

If Gaussian elimination for solving dense systems of linear equations is applied to the con uent Vandermonde-like systems, it requires $O(n^3)$ operations. However, the structure of the matrix V_c makes it possible to solve $V_c a = f$ and $V_c^T a = f$ with fewer arithmetic operations. A number of $O(n^2)$ fast algorithms for Vandermonde systems, Vandermonde-like systems, con uent Vandermonde systems, and con uent Vandermonde-like systems are presented in [5, 6, 7, 8, 9, 13, 15, 17, 20, 22, 23, 24, 25, 28, 30]. Some fast algorithms for inversion are available (see, for example, Traub [5] for Vandermonde matrices, Gohberg and Olshevsky [29] for Chebyshev–Vandermonde matrices, and Calvetti and Reichel [27] for Vandermonde-like matrices). The present paper describes a new fast algorithm for computing the elements of the inverse of

*Received by the editors July 26, 1996; accepted for publication (in revised form) by D. Calvetti September 8, 1997; published electronically April 2, 1998.
http://www.siam.org/journals/simax/19-3/30749.html
†Department of Mathematics, Northwestern Polytechnical University, Xian, People's Republic of China (shanlu@nwpu.edu.cn).
‡ Department of Mathematics, Xi'an Jiaotong University, Shaanxi Province, People's Republic of China.

an n × n con uent Vandermonde-like matrix V_c in $O(n^2)$ arithmetic operations if $n \gg \max_i n_i$. Our algorithm generalizes a scheme presented by Calvetti and Reichel [27] for fast inversion of Vandermonde-like matrices. It may be convenient to apply our scheme when many con uent Vandermonde-like systems have to be solved with the same matrix and different right-hand side vectors.

The paper is organized as follows. Section 2 describes our computational scheme. Computed examples in section 3 illustrate that our fast algorithm generally yields a more accurate approximation of the inverse than Gaussian elimination with partial pivoting when the nodes are ordered suitably.

2. Algorithm.

Set

$$m_1 = 0, \quad m_i = \sum_{k=1}^{i-1} n_k, \quad i = 1, 2, \ldots, p \tag{2.1}$$

and

$$\pi(x) = \prod_{k=1}^{p} (x-\alpha_k)^{n_k}, \quad L_i(x) = \prod_{\substack{k=1 \\ k\neq i}}^{p} (x-\alpha_k)^{n_k} = \frac{\pi(x)}{(x-\alpha_i)^{n_i}}, \quad i = 1, 2, \ldots, p. \tag{2.2}$$

Let each of the polynomials $P_{i,k}(x)$ be of degree $n_i - 1$ and satisfy

$$\frac{d^s}{dx^s}(P_{i,k}(x)L_i(x))\bigg|_{x=\alpha_i} = \delta_{ks}, \qquad k, s = 0, 1, \ldots, n_i - 1; \quad i = 1, 2, \ldots, p, \tag{2.3}$$

where δ_{ks} denotes the Kronecker δ-function.

If we obtain the representations of the polynomials $P_{i,k}(x)L_i(x)$ in terms of the basis $\{p_k(x)\}_{k=0}^{n-1}$,

$$P_{i,k}(x)L_i(x) = \sum_{t=1}^{n} u_{it}^{(k)} p_{t-1}(x), \qquad k = 0, 1, \ldots, n_i - 1; \quad i = 1, 2, \ldots, p, \tag{2.4}$$

and note that

$$\frac{d^s}{dx^s}(P_{i,k}(x)L_i(x))\bigg|_{x=\alpha_j} = 0, \qquad s = 0, 1, \ldots, n_j - 1; \quad j \neq i; \tag{2.5}$$

the inverse of V_c is, from (2.3), (2.5), and (2.4),

$$V_c^{-1} = \begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix}, \tag{2.6}$$

where

$$U_i = \begin{bmatrix} u_{i1}^{(0)} & \cdots & u_{in}^{(0)} \\ \cdots & \cdots & \cdots \\ u_{i1}^{(n_i-1)} & \cdots & u_{in}^{(n_i-1)} \end{bmatrix}, \qquad i = 1, 2, \ldots, p.$$

We describe a five-stage algorithm for computing the elements of V_c^{-1} as follows.

Stage 1. Express the polynomial $\pi(x)$ defined by (2.2) in terms of the polynomials $p_k(x)$, i.e.,

$$\pi(x) = \sum_{k=1}^{n+1} b_k p_{k-1}(x). \tag{2.7}$$

Let

$$\pi_0(x) = 1, \qquad \pi_j(x) = (x-\alpha_i)^{s+1} \prod_{k=1}^{i-1} (x-\alpha_k)^{n_k} = \sum_{k=1}^{j+1} b_k^{(j)} p_{k-1}(x), \tag{2.8}$$

where $j = m_i + s + 1$; $s = 0, 1, \ldots, n_i - 1$; $i = 1, 2, \ldots, p$. Substituting (2.8) into the recursion formula $\pi_j(x) = (x-\alpha_i)\pi_{j-1}(x)$ and applying (1.1),

$$(x-\beta_1)p_0(x) = \frac{1}{\theta_1} p_1(x), \quad (x-\beta_k)p_{k-1}(x) = \frac{1}{\theta_k} p_k(x) + \frac{\gamma_k}{\theta_k} p_{k-2}(x), \quad (k \geq 2) \tag{2.9}$$

yields

$$\begin{aligned}
\sum_{k=1}^{j+1} b_k^{(j)} p_{k-1}(x) &= \sum_{k=1}^{j} b_k^{(j-1)} (x-\beta_k) p_{k-1}(x) + \sum_{k=1}^{j} b_k^{(j-1)} (\beta_k - \alpha_i) p_{k-1}(x) \\
&= \left[(\beta_1 - \alpha_i) b_1^{(j-1)} + \frac{\gamma_2}{\theta_2}\, b_2^{(j-1)} \right] p_0(x) \\
&\quad + \sum_{k=2}^{j-1} \left[\frac{1}{\theta_k} b_{k-1}^{(j-1)} + (\beta_k - \alpha_i) b_k^{(j-1)} + \frac{\gamma_{k+1}}{\theta_{k+1}} b_{k+1}^{(j-1)} \right] p_{k-1}(x) \\
&\quad + \left[\frac{1}{\theta_{j-1}} b_{j-1}^{(j-1)} + (\beta_i - \alpha_i) b_j^{(j-1)} \right] p_{j-1}(x) + \frac{1}{\theta_j} b_j^{(j-1)} p_j(x),
\end{aligned}$$

which gives the recurrence relations.

ALGORITHM A.

$b_1^{(0)} = 1$
for $i = 1$ *to* p
  for $j = m_i + 1$ *to* $m_i + n_i$
    for $k = 1$ *to* $j + 1$

$$b_k^{(j)} = \frac{1}{\theta_{k-1}} b_{k-1}^{(j-1)} + (\beta_k - \alpha_i) b_k^{(j-1)} + \frac{\gamma_{k+1}}{\theta_{k+1}} b_{k+1}^{(j-1)}$$

    endfor k
  endfor j
endfor i,

where $\theta_0 = 1, b_0^{(j-1)} = 0$, *and* $b_k^{(j-1)} = 0$ *for* $k \geq j+1$. *Therefore*

$$b_k = b_k^{(n+1)}, \qquad k = 0, 1, \ldots, n+1.$$

Stage 2. Express the polynomial $L_i(x)$ defined by (2.2) in terms of the polynomials $p_k(x)$, i.e.,

$$L_i(x) = \sum_{k=1}^{n-n_i+1} c_{ik} p_{k-1}(x), \qquad i = 1, 2, \ldots, p. \tag{2.10}$$

Let

$$\frac{\pi(x)}{(x-\alpha_i)^j} = \sum_{k=1}^{n-j+1} c_{ik}^{(j)} p_{k-1}(x), \qquad j = 0, 1, \ldots, n_i; \quad i = 1, 2, \ldots, p. \tag{2.11}$$

Substituting (2.11) into the recursion formula

$$\frac{\pi(x)}{(x-\alpha_i)^{j-1}} = (x-\alpha_i)\frac{\pi(x)}{(x-\alpha_i)^j},$$

and applying (2.9) yields

$$\begin{aligned}\sum_{k=1}^{n-j+2} c_{ik}^{(j-1)} p_{k-1}(x) = (x-\alpha_i)\sum_{k=1}^{n-j+1} c_{ik}^{(j)} p_{k-1}(x) &= \left[(\beta_1-\alpha_i)c_{i1}^{(j)} + \frac{\gamma_2}{\theta_2}c_{i2}^{(j)}\right] p_0(x) \\ &+ \sum_{k=2}^{n-j}\left[\frac{1}{\theta_k}c_{i,k-1}^{(j)} + (\beta_k-\alpha_i)c_{ik}^{(j)} + \frac{\gamma_{k+1}}{\theta_{k+1}}c_{i,k+1}^{(j)}\right] p_{k-1}(x) \\ &+ \left[\frac{1}{\theta_{n-j}}c_{i,n-j}^{(j)} + (\beta_{n-j+1}-\alpha_i)c_{i,n-j+1}^{(j)}\right] p_{n-j}(x) + \frac{1}{\theta_{n-j+1}}c_{i,n-j+1}^{(j)} p_{n-j+1}(x),\end{aligned}$$

which gives the recurrence relations.

ALGORITHM B.

For $i = 1$ *to* p
  for $k = 1$ *to* $n+1$
    $c_{ik}^{(0)} = b_k$
  endfor k
  for $j = 1$ *to* n_i
    for $k = n-j+2$ *to* 2 *step* -1

$$c_{i,k-1}^{(j)} = \theta_{k-1}\left[c_{i,k}^{(j-1)} - (\beta_k-\alpha_i)c_{i,k}^{(j)} - \frac{\gamma_{k+1}}{\theta_{k+1}}c_{i,k+1}^{(j)}\right]$$

    endfor k
  endfor j
endfor i,

where $c_{ik}^{(j)} = 0$ *for* $k > n-j+1$. *Thus, from* (2.2),

$$c_{ik} = c_{ik}^{(n_i)}, \qquad k = 1, 2, \ldots, n-n_i+1; \quad i = 1, 2, \ldots, p.$$

Stage 3. Expand $L_i(x)$ in Taylor series at α_i, i.e.,

$$L_i(x) = d_{i1} + d_{i2}(x-\alpha_i) + \cdots + d_{i,n_i}(x-\alpha_i)^{n_i-1} + O((x-\alpha_i)^{n_i}). \tag{2.12}$$

We define

$$L_i^{(1)}(x) = L_i(x), \quad L_i^{(j)}(x) = d_{ij} + (x-\alpha_i)L_i^{(j+1)}(x), \qquad j = 1, \ldots, n_i. \tag{2.13}$$

Let

$$L_i^{(j)}(x) = \sum_{k=j}^{n-n_i+1} d_{ik}^{(j)} p_{k-j}(x). \tag{2.14}$$

Substituting (2.14) into (2.13), and applying (2.9), yields the following.

ALGORITHM C.

For $i = 1$ *to* p
  for $k = 1$ *to* $n - n_i + 1$
    $d_{ik}^{(1)} = c_{ik}$
  endfor k
  for $j = 1$ *to* n_i
    for $k = n - n_i + 1$ *to* $j + 1$ *step* -1

$$d_{i,k}^{(j+1)} = \theta_{k-j}\left[d_{i,k}^{(j)} - (\beta_{k-j+1} - \alpha_i)d_{i,k+1}^{(j+1)} - \frac{\gamma_{k-j+2}}{\theta_{k-j+2}}d_{i,k+2}^{(j+1)}\right]$$

    endfor k
    $d_{ij} = d_{ij}^{(j)} - (\beta_1 - \alpha_i)d_{i,j+1}^{(j+1)} - \dfrac{\gamma_2}{\theta_2}d_{i,j+2}^{(j+1)}$
  endfor j
endfor i,

where $d_{ik}^{(j+1)} = 0$ *for* $k > n - n_i + 1$.

Stage 4. Compute the polynomial $P_{i,k}(x)$ which is of degree $n_i - 1$ and satisfies (2.3).

Let

(2.15)

$$P_{i,k}(x) = p_{i1}^{(k)} + p_{i2}^{(k)}(x-\alpha_i) + \cdots + p_{i,n_i}^{(k)}(x-\alpha_i)^{n_i-1}, \quad k = 0, 1, \ldots, n_i - 1; \;\; i = 1, \ldots, p.$$

Substituting $x = \alpha_i$ into the formula

$$\frac{d^s}{dx^s}(P_{i,k}(x)L_i(x)) = \frac{d^s}{dx^s}(P_{i,k}(x))L_i(x) + s\frac{d^{s-1}}{dx^{s-1}}(P_{i,k}(x))\frac{d}{dx}(L_i(x)) + \frac{s(s-1)}{2}\frac{d^{s-2}}{dx^{s-2}}(P_{i,k}(x))\frac{d^2}{dx^2}(L_i(x)) + \cdots + P_{i,k}(x)\frac{d^s}{dx^s}(L_i(x)),$$

we have, from (2.15), (2.12), and (2.5),

$$\begin{bmatrix} d_{i1} & & & \\ d_{i2} & d_{i1} & & \\ \vdots & \ddots & \ddots & \\ d_{i,n_i} & \cdots & d_{i2} & d_{i1} \end{bmatrix} \begin{bmatrix} p_{i1}^{(0)} & p_{i1}^{(1)} & \cdots & p_{i1}^{(n_i-1)} \\ p_{i2}^{(0)} & p_{i2}^{(1)} & \cdots & p_{i2}^{(n_i-1)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{i,n_i}^{(0)} & p_{i,n_i}^{(1)} & \cdots & p_{i,n_i}^{(n_i-1)} \end{bmatrix} = \begin{bmatrix} 1 & & & \\ & \frac{1}{1!} & & \\ & & \ddots & \\ & & & \frac{1}{(n_i-1)!} \end{bmatrix},$$

where the coefficient matrix is a triangular Toeplitz matrix. Hence, we get the following.

ALGORITHM D.

For $i = 1$ *to* p
  $p_{i1}^{(0)} = \dfrac{1}{d_{i1}}$
  for $k = 2$ *to* n_i

$$p_{ik}^{(0)} = -\frac{1}{d_{i1}}\left(\sum_{t=1}^{k-1} d_{i,k-t+1}p_{it}^{(0)}\right)$$

endfor k
for $j = 1$ *to* $n_i - 1$
for $k = 1$ *to* n_i

$$p_{ik}^{(j)} = \begin{cases} \dfrac{1}{j}\, p_{i,k-1}^{(j-1)}, & k > j \\ 0, & k \le j \end{cases}$$

endfor k
endfor j
endfor i.

Stage 5. Compute the representations (2.4) of the polynomials $P_{i,k}(x)L_i(x)$ in terms *of the basis* $\{p_k(x)\}_{k=0}^{n-1}$.

We define

$$W_{i,n_i}^{(k)}(x) = p_{i,n_i}^{(k)}, \quad W_{ij}^{(k)}(x) = p_{ij}^{(k)} + (x - \alpha_i)W_{i,j+1}^{(k)}(x), \qquad j = n_i - 1, \dots, 1,$$

from which $W_{i,1}^{(k)}(x) = P_{i,k}(x)$. Let

$$W_{ij}^{(k)}(x)L_i(x) = \sum_{t=1}^{n-j+1} w_{i,t}^{(k,j)} p_{t-1}(x), \qquad j = 1, \dots, n_i; \quad i = 1, 2, \dots, p. \tag{2.16}$$

Substituting (2.16) into the recursion formula

$$\begin{cases} W_{i,n_i}^{(k)}(x)L_i(x) = p_{i,n_i}^{(k)} L_i(x), \\ W_{ij}^{(k)}(x)L_i(x) = p_{ij}^{(k)} L_i(x) + (x - \alpha_i)W_{i,j+1}^{(k)}(x)L_i(x), \qquad j = n_i - 1, \dots, 1, \end{cases}$$

and applying (2.10) and (2.9) yields the following.

ALGORITHM E.
For $i = 1$ *to* p
for $k = 0$ *to* $n_i - 1$
for $j = n_i$ *to* 1 *step* -1
for $t = 1$ *to* $n - j + 1$

$$w_{it}^{(k,j)} = p_{ij}^{(k)} c_{it} + \frac{1}{\theta_{t-1}}\, w_{i,t-1}^{(k,j+1)} + (\beta_t - \alpha_i) w_{it}^{(k,j+1)} + \frac{\gamma_{t+1}}{\theta_{t+1}} w_{i,t+1}^{(k,j+1)}$$

endfor t
endfor j
endfor k
endfor i,

where $w_{i0}^{(k,j)} = 0$, $w_{it}^{(k,j)} = 0 (j \le n_i, t > n - j + 1$ *or* $j > n_i)$, *and* $c_{it} = 0 (t > n - n_i + 1)$.

Finally, the elements of the inverse (2.6) of V_c are

$$u_{it}^{(k)} = w_{it}^{(k,1)}, \qquad t = 1, 2, \dots, n; \quad k = 0, 1, \dots, n_i - 1; \quad i = 1, 2, \dots, p.$$

Assuming that the values γ_j/θ_j are given, then the operation count of the algorithm A $\sim$ E can be bounded by

$$8n \sum_{i=1}^{p} n_i^2 - 4 \sum_{i=1}^{p} n_i^3 + 13n^2 - \frac{1}{2} \sum_{i=1}^{p} n_i^2 + \frac{27}{2}\, n + p,$$

which shows that the algorithm requires $O(n^2)$ arithmetic operations if $n \gg \max_i n_i$.

Of course, for $n_1 = \cdots = n_p = 1$, the five-stage algorithm reduces to the algorithm in [27].

If $\max_i n_i$ is large, then the arithmetic operations of the algorithm jumps to $O(n^3)$. For inversion of con uent Vandermonde-like matrices, an $O(n^2)$ algorithm was already specified in [31].

3. Computed examples. We present some experiments that illustrate the numerical behavior of the five-stage algorithm when the nodes are ordered in increasing, decreasing, and Leja orderings, which satisfy

$$|\alpha_1| = \max_{1\le s\le p} |\alpha_s|, \quad \prod_{j=1}^{k-1} |\alpha_k - \alpha_j| = \max_{k\le s\le p} \prod_{j=1}^{k-1} |\alpha_s - \alpha_j|, \quad 2 \le k \le p.$$

In order to compare the accuracy of the inverse V_c^{-1} computed by our algorithm with the three orderings of the nodes mentioned and computed by the Gaussian elimination with partial pivoting, we compute the Frobenius norm of the left residuals $\|V_c^{-1}V_c - I\|_F$. The computations were performed using FORTRAN 77 on a PC–AT compatible machine with double precision arithmetic.

Example 1. We use clustered nodes on $[-1, 1]$,

$$\alpha_j = -1 + 2\left(\frac{j-1}{p-1}\right)^2, \qquad j = 1, 2, \ldots, p \tag{2.17}$$

and the positive integers

$$n_j = \begin{cases} 5 - j, & j = 1, 2, 3, 4, \\ n_{j-4}, & j > 4. \end{cases} \tag{2.18}$$

The polynomials $p_k(x)$ are monomials, i.e., $p_k(x) = x^k$, $k = 0, 1, \ldots, n-1$. The results are given in Table 1.

Example 2. We use the clustered nodes (2.17) and the positive integers n_j (2.18). The orthogonal polynomials $p_k(x)$ are Chebyshev polynomials

$$p_k(x) = T_k(x) = \cos(k \arccos x), \qquad k = 0, 1, \ldots, n-1.$$

The results are given in Table 2.

Example 3. We use extrema nodes of the Chebyshev polynomial $T_{p-1}(x)$,

$$\alpha_j = \cos\left(\frac{j-1}{p-1}\,\pi\right), \qquad j = 1, 2, \ldots, p \tag{2.19}$$

and the positive integers

$$n_j = 2, \qquad j = 1, 2, \ldots, p. \tag{2.20}$$

The polynomials $p_k(x)$ are the monomials. The results are given in Table 3.

Example 4. The nodes and the positive integers are given in (2.19) and (2.20), respectively. The orthogonal polynomials $p_k(x)$ are the Chebyshev polynomials. The results are given in Table 4.

Example 5. We use equidistant nodes on $[-1, 1]$,

$$\alpha_j = -1 + 2\frac{j-1}{p-1}, \qquad j = 1, 2, \ldots, p,$$

and the positive integers n_j (see (2.20)). The orthogonal polynomials $p_k(x)$ are the Chebyshev polynomials. The results are given in Table 5.

Table 1

		Fast algorithm			
p	n	increasing	decreasing	Leja ordering	Gauss
4	10	0.288E 12	0.403E 12	0.615E 12	0.167E 11
8	20	0.574E 04	0.254E 02	0.736E 04	0.108E 01
12	30	0.340E+06	0.725E+06	0.929E+04	0.299E+09

Table 2

		Fast algorithm			
p	n	increasing	decreasing	Leja ordering	Gauss
4	10	0.236E 12	0.180E 11	0.234E 12	0.133E 11
8	20	0.776E 03	0.113E 02	0.609E 07	0.813E 06
12	30	0.181E+04	0.262E+06	0.282E+03	0.312E+01

Table 3

		Fast algorithm			
p	n	increasing	decreasing	Leja ordering	Gauss
5	10	0.579E 12	0.136E 11	0.191E 12	0.674E 12
10	20	0.593E 06	0.134E 05	0.326E 08	0.666E 04
15	30	0.322E+00	0.177E+00	0.773E 04	0.957E+03
20	40	0.331E+04	0.210E+04	0.498E+00	

Table 4

		Fast algorithm			
p	n	increasing	decreasing	Leja ordering	Gauss
5	10	0.116E 11	0.833E 12	0.291E 13	0.472E 14
10	20	0.679E 06	0.496E 06	0.139E 11	0.127E 12
15	30	0.995E 01	0.944E 01	0.138E 10	0.161E 12
20	40	0.910E+03	0.171E+04	0.816E 10	0.474E 12
25	50			0.103E 09	0.102E 11
30	60			0.273E 09	0.231E 11
35	70			0.182E 08	0.296E 11
40	80			0.184E 08	0.518E 11
45	90			0.994E 08	0.595E 11

Table 5

		Fast algorithm			
p	n	increasing	decreasing	Leja ordering	Gauss
5	10	0.422E 14	0.422E 14	0.422E 14	0.638E 14
10	20	0.190E 06	0.190E 06	0.163E 10	0.225E 11
15	30	0.116E+01	0.116E+01	0.572E 07	0.320E 08
20	40			0.382E 04	0.175E 04
25	50			0.475E 01	0.232E+00
30	60			0.168E+02	0.373E+03

Our numerical experiments suggest that, in general, our fast algorithm with Leja ordering of the nodes is at least as highly accurate in the computed inverse of a con uent Vandermonde-like matrix as Gaussian elimination with partial pivoting.

Acknowledgments. We thank the referees for their valuable comments.

REFERENCES

[1] N. Macon and A. Spitzbart, *Inverses of Vandermonde matrices*, Amer. Math. Monthly, 65 (1958), pp. 95–100.
[2] W. Gautschi, *On inverses of Vandermonde and con uent Vandermonde matrices*, Numer. Math., 4 (1962), pp. 117–123.
[3] W. Gautschi, *On inverses of Vandermonde and con uent Vandermonde matrices* II, Numer. Math., 5 (1963), pp. 425–430.
[4] J. N. Lyness and C. B. Moler, *Van Der Monde systems and numerical differentiation*, Numer. Math., 8 (1966), pp. 458–464.
[5] J. F. Traub, *Associated polynomials and uniform methods for the solution of linear problems*, SIAM Rev., 8 (1966), pp. 277–301.
[6] C. Ballester and V. Pereyra, *On the construction of discrete approximations to linear differential expressions*, Math. Comp., 21 (1967), pp. 297–302.
[7] A. Björck and V. Pereyra, *Solution of Vandermonde systems of equations*, Math. Comp., 24 (1970), pp. 893–903.
[8] G. Galimberti and V. Pereyra, *Solving con uent Vandermonde systems of Hermite type*, Numer. Math., 18 (1971), pp. 44–60.
[9] A. Björck and T. Elfving, *Algorithms for con uent Vandermonde systems*, Numer. Math., 21 (1973), pp. 130–137.
[10] W. Gautschi, *Norm estimates for inverses of Vandermonde matrices*, Numer. Math., 23 (1975), pp. 337–347.
[11] H. Van de Vel, *Numerical treatment of a generalized Vandermonde system of equations*, Linear Algebra Appl., 17 (1977), pp. 149–179.
[12] W. Gautschi, *On inverses of Vandermonde and con uent Vandermonde matrices* III, Numer. Math., 29 (1978), pp. 445–450.
[13] W. P. Tang and G. H. Golub, *The block decomposition of a Vandermonde matrix and its applications*, BIT, 21 (1981), pp. 505–517.
[14] W. Gautschi, *The condition of Vandermonde-like matrices involving orthogonal polynomials*, Linear Algebra Appl., 52/53 (1983), pp. 293–300.
[15] I. Gohberg, T. Kailath, and I. Koltracht, *Efficient solution of linear systems of equations with recursive structure*, Linear Algebra Appl., 80 (1986), pp. 81–113.
[16] N. J. Higham, *Error analysis of the Björck–Pereyra algorithms for solving Vandermonde systems*, Numer. Math., 50 (1987), pp. 613–632.
[17] I. Gohberg, T. Kailath, I. Koltracht, and P. Lancaster, *Linear complexity parallel algorithms for linear systems of equations with recursive structure*, Linear Algebra Appl., 88/89 (1987), pp. 271–315.
[18] W. Gautschi and G. Inglese, *Lower bounds for the condition number of Vandermonde matrices*, Numer. Math., 52 (1988), pp. 241–250.
[19] L. Verde-Star, *Inverses of generalized Vandermonde matrices*, J. Math. Anal. Appl., 131 (1988), pp. 341–353.
[20] N. J. Higham, *Fast solution of Vandermonde-like systems involving orthogonal polynomials*, IMA J. Numer. Anal., 8 (1988), pp. 473–486.
[21] C. J. Demeure, *Fast QR factorization of Vandermonde matrices*, Linear Algebra Appl., 122/123/124 (1989), pp. 165–194.
[22] N. J. Higham, *Stability analysis of algorithms for solving con uent Vandermonde-like systems*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 23–41.
[23] L. Reichel and G. Opfer, *Chebyshev–Vandermonde systems*, Math. Comp., 57 (1991), pp. 703–721.
[24] J. Chun and T. Kailath, *Displacement structure for Hankel, Vandermonde, and related (derived) matrices*, Linear Algebra Appl., 151 (1991), pp. 199–227.

[25] D. Calvetti and L. Reichel, *A Chebyshev–Vandermonde solver*, Linear Algebra Appl., 172 (1992), pp. 219–229.
[26] T. Finck, G. Heinig, and K. Rost, *An inversion formula and fast algorithms for Cauchy–Vandermonde matrices*, Linear Algebra Appl., 183 (1993), pp. 179–191.
[27] D. Calvetti and L. Reichel, *Fast inversion of Vandermonde-like matrices involving orthogonal polynomials*, BIT, 33 (1993), pp. 471–484.
[28] Hao Lu, *Fast solution of con uent Vandermonde linear systems*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 1277–1289.
[29] I. Gohberg and V. Olshevsky, *Fast inversion of Chebyshev Vandermonde matrices*, Numer. Math., 67 (1994), pp. 71–92.
[30] Hao Lu, *Fast algorithms for con uent Vandermonde linear systems and generalized Trummer's problem*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 655–674.
[31] G. Heinig, W. Hoppe, and K. Rost, *Structured matrices in interpolation and approximation problems*, Wiss. Z. Tech. Univ. Karl Marx Stadt, 31 (1989), pp. 196–202.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 807–815, July 1998

014

GROWTH IN GAUSSIAN ELIMINATION, ORTHOGONAL MATRICES, AND THE 2-NORM*

JESSE L. BARLOW† AND HONGYUAN ZHA†

Abstract. It is shown that maximal growth for Gaussian elimination with partial pivoting as measured in the 2-norm is achieved by orthogonal matrices. A precise bound on that growth is given.

Key words. LU factorization, orthogonal invariance, triangular matrix

AMS subject classifications. 65F05, 65F35

PII. S0895479896309912

1. Introduction.

Consider Gaussian elimination with partial pivoting (GEPP) applied to a matrix $A \in \Re^{n\times n}$. The algorithm produces a decomposition of the form

$$A = PLU, \tag{1.1}$$

where $L \in \Re^{n\times n}$ is unit lower triangular, $U \in \Re^{n\times n}$ is upper triangular, and $P \in \Re^{n\times n}$ is a permutation matrix.

For simplicity, assume that the permutation matrix P is already known. GEPP generates a sequence of matrices $A_k = (a_{ij}^{(k)}), k = 0, \ldots, n-1$, defined inductively by

$$A_0 = P^T A, \tag{1.2}$$

$$A_k = (I - \mathbf{m}_k \mathbf{e}_k^T) A_{k-1}, \quad k = 1, \ldots, n-1, \tag{1.3}$$

where

$$\mathbf{m}_k = \frac{1}{a_{kk}^{(k-1)}} (\overbrace{0, \ldots, 0}^{k}, a_{k+1,k}^{(k-1)}, \ldots, a_{nk}^{(k-1)})^T. \tag{1.4}$$

It is easily shown that

$$L_k A_k = A_0, \quad k = 1, \ldots, n-1, \tag{1.5}$$

where

$$L_k = I + \sum_{i=1}^{k} \mathbf{m}_i \mathbf{e}_i^T, \quad k = 1, \ldots, n-1. \tag{1.6}$$

The permutation P is chosen so that

$$|a_{kk}^{(k-1)}| = \max_{k\le i\le n} |a_{ik}^{(k-1)}|. \tag{1.7}$$

*Received by the editors September 24, 1996; accepted for publication (in revised form) by N.J. Higham July 11, 1997; published electronically April 14, 1998. The research of the first author was supported by National Science Foundation grants CCR-9201612 and CCR-9424435. Part of this work was done while the first author was visiting the University of Manchester, Department of Mathematics, Manchester M13 9PL, United Kingdom.

http://www.siam.org/journals/simax/19-3/30971.html

†Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802-6106 (barlow@cse.psu.edu, zha@cse.psu.edu).

If we let

$$L = L_{n-1}, \quad U = A_{n-1},$$

we obtain the decomposition of A given by (1.1).

Equation (1.7) ensures that $L = (\ell_{ij})$, where $|\ell_{ij}| \leq 1$ for all i and j, and that the same holds for each L_k.

Wilkinson [7] showed that the stability of GEPP depends upon the growth factor

$$\zeta_A^{(k)} = \frac{\max_{(i,j)} |a_{ij}^{(k)}|}{\max_{(i,j)} |a_{ij}|}.$$

He also showed that

$$\max_{A \in \Re^{n \times n}} \zeta_A^{(k)} = 2^k$$

and

$$\zeta_n = \max_{A \in \Re^{n \times n}} \zeta_A = 2^{n-1}.$$

The decomposition (1.1) is stable for a particular A if

$$\zeta_A = \max_{1 \leq k \leq n-1} \zeta_A^{(k)}$$

is a modest value.

For our discussion, we define the growth factor in the 2-norm as

$$\rho_A^{(k)} = \frac{\|A_k\|_2}{\|A\|_2}. \tag{1.8}$$

Standard norm inequalities lead to the bounds

$$\frac{1}{n} \rho_A^{(k)} \leq \zeta_A^{(k)} \leq n \rho_A^{(k)}.$$

We also define the bounds

$$\rho_A = \max_{1 \leq k \leq n-1} \rho_A^{(k)}, \quad \rho_n^{(k)} = \max_{A \in \Re^{n \times n}} \rho_A^{(k)}, \quad \rho_n = \max_{1 \leq k \leq n-1} \rho_n^{(k)}. \tag{1.9}$$

We prove the following two results about $\rho_A^{(k)}$.

PROPOSITION 1.1. *Let $A \in \Re^{n \times n}$ be nonsingular and have the P-L-U factorization* (1.1) *by GEPP. Let A have the factorization*

$$A = QR, \tag{1.10}$$

where $Q \in \Re^{n \times n}$ is orthogonal and $R \in \Re^{n \times n}$ is upper triangular. Then GEPP on Q produces the P-L-U factorization

$$Q = PLU_Q, \tag{1.11}$$

where

$$U_Q = UR^{-1}. \tag{1.12}$$

If we let $\rho_A^{(k)}$ *and* $\rho_Q^{(k)}$ *be as defined by* (1.8), *then*

$$\rho_Q^{(k)} \geq \rho_A^{(k)}, \quad k = 1, \ldots, n-1. \tag{1.13}$$

Thus $\rho_Q \geq \rho_A$.

Proposition 1.1 states that orthogonal matrices are the worst case for growth in GEPP. The second proposition establishes a sharp bound for that growth.

PROPOSITION 1.2. *Let* $\rho_n^{(k)}$ *and* ρ_n *be defined by* (1.9). *Then*

$$\rho_n^{(k)} + O(\sqrt{n}) = \frac{1}{\sqrt{3}} \left((n-k+1/3)\right)^{1/2} 2^k, \tag{1.14}$$

and therefore

$$\rho_n + O(\sqrt{n}) = \frac{2^n}{3}. \tag{1.15}$$

Proposition 1.1 arose from work by the first author [1] on the error analysis of bidiagonal reduction. In that work, it was necessary to understand GEPP applied to orthogonal matrices. The two authors then proved Proposition 1.2 to give a precise value for worst-case growth.

In the next section, we prove Propositions 1.1 and 1.2. In section 3, we revisit Wilkinson's famous example of growth in GEPP and state our conclusions.

2. Proofs of the propositions. We now prove Propositions 1.1 and 1.2. The proof of Proposition 1.1 is quite short, and we give it first.

Proof of Proposition 1.1. Combining (1.1) and (1.10) and using the nonsingularity of A obtain

$$Q = AR^{-1} = PLUR^{-1}. \tag{2.1}$$

Since U and R^{-1} are upper triangular, so is $U_Q = UR^{-1}$; thus we have (1.11).

By a similar argument,

$$Q = PL_kQ_k, \tag{2.2}$$

where

$$Q_k = A_kR^{-1}$$

and L_k and A_k are given in (1.5) and (1.6). Since the first k columns of Q_k are zero below the diagonal, the uniqueness of the L-U decomposition [6, p. 121, Theorem 2.6] assures us that (2.2) is the kth stage of GEPP applied to Q. No row interchanges are needed since all entries of L_k below the diagonal will have an absolute value less than 1.

A straightforward use of norm inequalities yields

$$\rho_A^{(k)} = \frac{\|A_k\|_2}{\|A\|_2} = \frac{\|L_k^{-1}A_0\|_2}{\|A_0\|_2} \leq \|L_k^{-1}\|_2. \tag{2.3}$$

However,

$$\rho_Q = \frac{\|Q_k\|_2}{\|Q\|_2} = \|L_k^{-1}P^TQ\|_2 = \|L_k^{-1}\|_2, \tag{2.4}$$

where equality holds since P^TQ is orthogonal. Combining (2.3) and (2.4) yields (1.13). ∎

REMARK 2.1. *The above result does not extend to complete pivoting. The following example, generated using the MATLAB function randn, illustrates this fact. It is given to six fixed point digits. Let*

$$A = \begin{pmatrix} 1.164954 & 0.351607 & 0.059060 \\ 0.626839 & -0.696513 & 1.797072 \\ 0.075080 & 1.696142 & 0.264069 \end{pmatrix}.$$

Its orthogonal factor is

$$Q = \begin{pmatrix} -0.879196 & 0.152789 & -0.451298 \\ -0.473079 & -0.392581 & 0.788719 \\ -0.056663 & 0.906938 & 0.417437 \end{pmatrix}.$$

Gaussian elimination with complete pivoting applied to A obtains the row permutation $p_r = (2,3,1)^T$ and the column permutation $p_c = (3,2,1)^T$. The L and U factors are

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0.146944 & 1 & 0 \\ 0.032864 & 0.208229 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 1.797072 & -0.696513 & 0.626839 \\ 0 & 1.798491 & -0.017030 \\ 0 & 0 & 1.147899 \end{pmatrix}.$$

On the other hand, Gaussian elimination with complete pivoting applied to Q obtains the row permutation $p_r = (3,2,1)^T$ and the column permutation $p_c = (2,3,1)^T$. The L and U factors are

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -0.432864 & 1 & 0 \\ 0.168467 & -0.538081 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 0.906938 & 0.417437 & -0.056663 \\ 0 & 0.969412 & -0.497606 \\ 0 & 0 & -1.137403 \end{pmatrix}.$$

Three lemmas are necessary to prove Proposition 1.2. They concern the matrix $\tilde{L}_k = (\tilde{\ell}_{ij}^{(k)})$, defined by

$$\tilde{\ell}_{ij}^{(k)} = \begin{cases} 1 & \text{if } i = j, \\ -1 & \text{if } 1 \le j \le k \text{ and } j < i, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

The first lemma is a slightly different version of a bound for the norm of the inverse of a triangular matrix given in Higham [3, pp. 159–161, Theorems 8.11 and 8.13]. The proof is nearly identical to what is given there, so we omit it.

LEMMA 2.1. *Let $L_k = (\ell_{ij}^{(k)}) \in \Re^{n\times n}$ be a lower triangular matrix such that $\ell_{ii}^{(k)} = 1$, $|\ell_{ij}^{(k)}| \le 1$ for $j \le \min\{i-1, k\}$, and $\ell_{ij} = 0$ if $i > j > k$. Let $\tilde{L}_k$ be defined by* (2.5). *Then*

$$\|L_k^{-1}\|_2 \le \|\tilde{L}_k^{-1}\|_2. \tag{2.6}$$

The second lemma bounds $\|\tilde{L}_k^{-1}\|_2$. It is similar to a bound given by Faddeev, Kublanovskaya, and Faddeeva [2], whose proof is given by Lawson and Hanson [4, pp. 28–35].

LEMMA 2.2. *Let $\tilde{L}_k \in \Re^{n\times n}$ be defined by* (2.5) *and let $\rho_n^{(k)}$ be defined by* (1.9). *Then*

$$\rho_n^{(k)} = \|\tilde{L}_k^{-1}\|_2 \le \|\tilde{L}_k^{-1}\|_F = \left((n-k+1/3)\frac{4^k-1}{3} + n - \frac{1}{3}k\right)^{1/2}.$$

Proof. From Lemma 2.1, it follows that

$$\rho_n^{(k)} = \|\tilde{L}_k\|_2 \le \|\tilde{L}_k\|_F. \tag{2.7}$$

Define $\mathbf{w}_j, \mathbf{c}_j \in \Re^j, j = 1, \ldots, n$, by

$$\mathbf{w}_j = (1, 1, 2, \ldots, 2^{j-2})^T, \quad \mathbf{c}_j = (1, 1, \ldots, 1)^T. \tag{2.8}$$

Note that if $j \le k$, then

$$\tilde{L}_k^{-1}\mathbf{e}_j = \begin{matrix} {\scriptstyle j-1} \\ {\scriptstyle k-j+1} \\ {\scriptstyle n-k} \end{matrix} \begin{pmatrix} 0 \\ \mathbf{w}_{k-j+1} \\ 2^{k-j}\mathbf{c}_{n-k} \end{pmatrix}.$$

If $j > k$, then

$$\tilde{L}_k^{-1}\mathbf{e}_j = \mathbf{e}_j.$$

Using the summing formula for a geometric series yields

$$\|\tilde{L}_k^{-1}\mathbf{e}_j\|_2^2 = \begin{cases} \frac{4^{k-j}+2}{3} + (n-k)4^{k-j} & \text{if } j \le k, \\ 1 & \text{if } j > k. \end{cases}$$

If we then use the summing formula for a geometric series again and take square roots, we obtain the bound claim. ∎

The following corollary gives a class of matrices that always achieve the optimal growth in GEPP. Its proof is obvious.

COROLLARY 2.3. *Let $A \in \Re^{n\times n}$ and let $\tilde{L}_{n-1}$ be given by* (2.5). *Let*

$$A = P\tilde{L}_{n-1}U$$

be the factorization of A by GEPP. If Q is the orthogonal factor of A, as defined in Proposition 1.1, *then*

$$\rho_Q^{(k)} = \rho_n^{(k)}, \quad k = 1, \ldots, n-1.$$

In the next lemma, for a matrix $B \in \Re^{m\times n}$ we define $\sigma_i(B)$ to be the ith singular value of B (in decreasing order) for $i = 1, \ldots, n$.

The upper bound from Lemma 2.2 is very tight. The matrix $\tilde{L}_k$ always has exactly one small singular value.

LEMMA 2.4. *Let $\tilde{L}_k$ be as in Lemma* 2.1. *Assume that $n \ge 2$. If $k < n-1$, $\sigma_{n-1}(\tilde{L}_k) = 1$; otherwise $\sigma_{n-1}(\tilde{L}_k) \ge \sqrt{2}$.*

Proof. First consider the case $k = n-1$. Let

$$\tilde{L}_{n-1} = \begin{pmatrix} F_{n-1} & \mathbf{e}_n \end{pmatrix}. \tag{2.9}$$

The Cauchy interlace theorem [5, p. 186, Theorem 10-1-2], when applied to $\tilde{L}_{n-1}^T\tilde{L}_{n-1}$, states that

$$\sigma_{n-1}(\tilde{L}_{n-1}) \ge \sigma_{n-1}(F_{n-1}).$$

We now show that $\sigma_{n-1}(F_{n-1}) = \sqrt{2}$.

Note that

$$F_{n-1}^T F_{n-1} = \begin{matrix} \\ {\scriptstyle n-2} \\ {\scriptstyle 1} \end{matrix} \begin{matrix} \begin{matrix} {\scriptstyle n-2} & {\scriptstyle 1} \end{matrix} \\ \begin{pmatrix} B_{n-2} & 0 \\ 0 & 2 \end{pmatrix} \end{matrix},$$

where B_{n-2} is empty if $n = 2$ and

$$B_{n-2} = \begin{pmatrix} n & n-3 & n-4 & \dots & 1 \\ n-3 & n-1 & n-4 & \dots & 1 \\ n-4 & n-5 & n-2 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 3 \end{pmatrix}.$$

It is obvious that $\sqrt{2}$ is a singular value of F_{n-1}. If $n = 2$, it is clearly the smallest positive one, and we are done.

For $n \geq 3$, we now show that $B_{n-2} - 2I_{n-2}$ is positive definite. We have

$$\hat{B}_{n-2} = B_{n-2} - 2I_{n-2} = \begin{pmatrix} n-2 & n-3 & n-4 & \dots & 1 \\ n-3 & n-3 & n-4 & \dots & 1 \\ n-4 & n-4 & n-4 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix}.$$

The matrix $\hat{B}_{n-2}$ can be factored into

$$\hat{B}_{n-2} = G_{n-2}^T G_{n-2},$$

where

$$G_{n-2} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}.$$

Thus $\hat{B}_{n-2}$ is positive definite, $\sqrt{2}$ must be the smallest singular value of F_{n-1}, and $\sigma_{n-1}(\tilde{L}_{n-1}) \geq \sqrt{2}$.

To consider $k < n-1$, we need some helpful notation. For any fixed n and $1 \leq k \leq n-1$, define

$$F_{k,n} = \tilde{L}_k$$

for $k = 1, \dots, n-1$.

We have that

$$F_{k,n} = \begin{matrix} \\ {\scriptstyle k} \\ {\scriptstyle n-k} \end{matrix} \begin{matrix} \begin{matrix} {\scriptstyle k} & {\scriptstyle n-k} \end{matrix} \\ \begin{pmatrix} F_{k-1,k} & 0 \\ J_{n-k} & I_{n-k} \end{pmatrix} \end{matrix},$$

where J_{n-k} is an $(n-k) \times k$ matrix with each entry -1. Clearly, J_{n-k} is a rank-1 matrix. Thus

$$\dim(\text{null}(J_{n-k}^T)) = n - k - 1.$$

If $\mathbf{z} \in \text{null}(J_{n-k}^T)$, then $\hat{\mathbf{z}} = (\overbrace{0, \ldots, 0}^{k}, \mathbf{z})^T$ satisfies

$$F_{k,n} F_{k,n}^T \hat{\mathbf{z}} = \hat{\mathbf{z}}.$$

Thus $n-k-1$ singular values of $F_{k,n}$ are equal to 1. From our previous argument, $\sigma_{k-1}(F_{k,k-1}) \geq \sqrt{2}$, so again, applying the Cauchy interlace theorem to $F_{k,n}^T F_{k,n}$ yields

$$\sigma_{k-1}(F_{k,n}) \geq \sigma_{k-1}(F_{k,k-1}) \geq \sqrt{2}.$$

Therefore we have

$$\sigma_k(F_{k,n}) = \cdots = \sigma_{n-1}(F_{k,n}) = 1. \qquad \square$$

We can now prove Proposition 1.2. *Proof of Proposition* 1.2. We have that

$$\rho_n^{(k)} = \|\tilde{L}_k^{-1}\|_2 = \sigma_n(\tilde{L}_k)^{-1}.$$

For simplicity, let $\sigma_i = \sigma_i(\tilde{L}_k)$.

We also have that

$$\sigma_n^{-2} = \|\tilde{L}_k^{-1}\|_F^2 - \sum_{i=1}^{n-1} \sigma_i^{-2}.$$

From Lemma 2.4, we have that $\sigma_{n-1} \geq 1$, thus $\sigma_{n-1}^{-2} \leq 1$, so

$$\sigma_n^{-2} \geq \|\tilde{L}_k^{-1}\|_F^2 - (n-1)$$

$$= (n-k+1/3)\frac{4^k}{3} - \frac{n-2}{3}.$$

Taking square roots and using the reverse triangle inequality yield

$$\sigma_n^{-1} \geq \frac{1}{\sqrt{3}}((n-k+1/3)^{1/2}\, 2^k - (n-2)^{1/2}) = \frac{1}{\sqrt{3}}\,(n-k+1/3)^{1/2}\, 2^k - O(\sqrt{n}).$$

This establishes (1.14). Equation (1.15) follows from maximizing (1.14) over all $k = 1, \ldots, n-1$. $\square$

REMARK 2.2. *An approximate largest right singular vector of* $\tilde{L}_k^{-1}$ *is*

$$\mathbf{x} = (2^{k-1}, 2^{k-2}, \ldots, 2, 1, 1, \ldots, 1)^T.$$

For this vector $\|\tilde{L}_k^{-1}\mathbf{x}\|_2 / \|\mathbf{x}\|_2 \approx \rho_n^{(k)} + o(2^k)$.

REMARK 2.3. *If we were to use the measure*

$$\gamma_A^{(k)} = \frac{\|A_k\|_F}{\|A\|_2},$$

the results of Lemma 2.2 *could be used to show that*

$$\gamma_n^{(k)} = \max_{A \in \Re^{n \times n}} \gamma_A^{(k)} = \left((n-k+1/3)\frac{4^k - 1}{3} + n - \frac{1}{3}k\right)^{1/2}.$$

Moreover, in the proof of Proposition 1.1*, orthogonal invariance yields*

$$\|Q_k\|_F = \|L_k^{-1}\|_F \|Q\|_2.$$

Thus orthogonal matrices would also maximize this measure.

However, orthogonal matrices do not maximize the measure

$$\tilde{\gamma}_A^{(k)} = \frac{\|A_k\|_F}{\|A\|_F}.$$

We note that

$$\tilde{\gamma}_A^{(k)} \le \|L_k^{-1}\|_2 \le \rho_n^{(k)}.$$

Therefore, from Remark 2.2*, if we let F_n be the matrix defined by* (2.9) *and let $\mathbf{x}_n = (1, 2^{-1}, \ldots, 2^{-(n-1)}, 2^{-(n-1)})^T$, then the upper bound for $\tilde{\gamma}_A^{(n)}$ can be achieved asymptotically by the sequence of matrices*

$$A^{(n)} = \begin{pmatrix} n^{-2}F_n & \mathbf{x}_n \end{pmatrix}$$

as $n \to \infty$.

3. Wilkinson's example revisited and the conclusion. Let $A = (a_{ij}) \in \Re^{25\times 25}$ be the matrix

$$a_{ij} = \begin{cases} 0 & \text{if } i < j < n, \\ 1 & \text{if } i = j \text{ or } j = n, \\ -1 & \text{if } i > j. \end{cases}$$

This is a famous example due to Wilkinson [8, section 4.26]. We then let the *P-L-U* factorization by GEPP be given by (1.1) and the orthogonal factorization of A be given by (1.10).

With MATLAB one finds that

$$\rho_A = 1.2370 \times 10^6.$$

However, the growth factor ρ_Q for the orthogonal factor Q in (1.10) is

$$\rho_Q = 1.1185 \times 10^7 > \rho_A.$$

We note that

$$\|L^{-1}\|_F = \left(4^{25} + 6 \times 25 - 1\right)^{1/2} = 1.1185 \times 10^7,$$

so it matches ρ_Q to five significant digits.

This example points out that Wilkinson's matrix does not produce the largest possible growth in the 2-norm, but as Corollary 2.3 states, its orthogonal factor does.

The MATLAB 4.2c routine lu(X) pivoted when performed on the Q from this example, but still produces a very large growth factor. Thus we had to factor Q using a Gaussian elimination routine that did not pivot to get these results.

As we have shown, orthogonal matrices are the worst case for 2-norm growth in GEPP.

Acknowledgments. Nick Higham and two thorough referees made very helpful comments. In particular, Remark 2.1 came from one of Nick Higham's questions, and Remark 2.3 came from a referee's question. A second referee recommended notational changes that clarified the presentation.

REFERENCES

[1] J. Barlow, *More Accurate Bidiagonal Reduction for Computing the Singular Value Decomposition*, manuscript, 1997.
[2] D. Faddeev, V. Kublanovskaya, and V. Faddeeva, *Solution of linear algebraic systems with rectangular matrices*, Proc. Steklov Inst. Math., 96 (1968), pp. 93–111.
[3] N. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.
[4] C. Lawson and R. Hanson, *Solving Least Squares Problems*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[5] B. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[6] G. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
[7] J. Wilkinson, *Error analysis of direct methods of matrix inversion*, J. Assoc. Comput. Mach., 8 (1961), pp. 281–330.
[8] J. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 816–832, July 1998

015

ROBUST ORDERING OF SPARSE MATRICES USING MULTISECTION*

CLEVE ASHCRAFT† AND JOSEPH W.H. LIU‡

Abstract. In this paper we provide a robust reordering scheme for sparse matrices. The scheme relies on the notion of *multisection*, a generalization of bisection. The reordering strategy is demonstrated to have consistently good performance in terms of fill reduction when compared with multiple minimum degree and generalized nested dissection. Experimental results show that by using multisection, we obtain an ordering which is consistently as good as or better than both for a wide spectrum of sparse problems.

Key words. ordering algorithms, minimum degree algorithm, nested dissection

AMS subject classifications. 65F05, 65F50, 68R10

PII. S0895479896299081

1. Introduction.

It is well recognized that finding a fill-reducing ordering is crucial to the success of the numerical solution of sparse linear systems. For symmetric positive-definite systems, the minimum degree [47] and the nested dissection [15] orderings are perhaps the most popular ordering schemes. They represent two opposite approaches to the ordering problem. Minimum degree is a "bottom-up" approach that uses local information, while nested dissection is a "top-down" approach that primarily uses global information. However, the two methods share a common undesirable characteristic. Both schemes produce generally good orderings, but the ordering quality is not uniformly good.

The main contribution of this paper is to introduce a robust ordering scheme that gives good quality orderings consistently, near equal or better than minimum degree and nested dissection. The basic tool is a *multisector*, a generalization of a bisector. A multisector is a subset of vertices whose removal subdivides the graph into two or more components. We call the resulting partition a *domain decomposition*. The whole ordering process has two *independent* phases: the ordering of the vertices in the components (the domains) and the ordering of the multisector (the interface).

An outline of this paper is as follows. In section 2, we provide evidence on the inconsistent ordering quality of the minimum degree and nested dissection schemes. The multisection ordering scheme is described in section 3, where the notions of domain decomposition, multisector, and multisection ordering are introduced. The quality of the resulting ordering depends on three factors: the domain decomposition, the ordering method for the domains, and the ordering method for the multisector.

In section 4, we consider numerical experiments of the multisection ordering approach on regular grids. We show that multisection gives an ordering quality close to the optimal nested dissection ordering [15] for square and cubic grids, and local

*Received by the editors February 16, 1996; accepted for publication (in revised form) by D. Calvetti May 19, 1997; published electronically April 28, 1998.

http://www.siam.org/journals/simax/19-3/29908.html

†Boeing Information and Support Services, P. O. Box 24346, Mail Stop 7L-22, Seattle, WA 98124 (cleve.ashcraft@boeing.com). This research was supported in part by ARPA contract DABT63-95-C-0122.

‡Department of Computer Science, York University, North York, Ontario M3J 1P3, Canada (joseph@cs.york.ca). This research was supported in part by the Natural Sciences and Engineering Research Council of Canada under grant A5509 and in part by ARPA contract DABT63-95-C-0122.

nested dissection ordering [9] for grids with large aspect ratios. Section 5 evaluates multisection on some structural analysis matrices from the Harwell–Boeing collection [12]. We use an incomplete nested dissection scheme to determine a multisector and its associated domain decomposition. Section 6 contains our concluding remarks.

2. Minimum degree and nested dissection orderings.

2.1. General overview. The *minimum degree ordering* algorithm is a symmetric version of the Markowitz scheme [39]. It was first described and used by Tinney and Walker [47]. The basic minimum degree algorithm can be best described in terms of elimination graphs [44]. Let G be the graph associated with a given sparse matrix. The scheme selects a vertex v of minimum degree in G. This vertex is numbered next in the ordering and is eliminated from the graph G to form its elimination graph G_v. The graph G is then replaced by this elimination graph G_v, where the selection/elimination process is repeated. Many important enhancements have been made to the implementation of the minimum degree ordering; the survey paper [14] contains a comprehensive account of such enhancements.

To choose a vertex to eliminate, the minimum degree algorithm uses the degree of the vertex, which is a *local* graph property. If we view the ordering as the construction of the elimination tree,[1] the tree is formed from bottom-up. This means vertices associated with the bottom part of the tree get their ordering assignments first.

The minimum degree ordering has been generally recommended as a general purpose fill-reducing reordering scheme. Its wide acceptance is largely due to its effectiveness in reducing fill and its efficient implementation. However, since the scheme uses only local information, it produces only adequate orderings for large problems. There is room for improvement.

Another popular fill-reducing ordering is the *nested dissection ordering* [15] and its generalizations. In contrast to the minimum degree algorithm, nested dissection is a top-down scheme. It finds a *separator*, a subset of vertices whose removal renders the remaining graph disconnected, and numbers the vertices in the separator last. The process is then repeated recursively on each of the resulting components. Nested dissection constructs the elimination tree from the root down.

In [15] and [27] it is showed that nested dissection on grid problems will give optimal orderings (to within a constant factor) with respect to the number of factor nonzeros and factor operation counts. For general graph problems, the recursive approach of finding and ordering separators is often referred to as *generalized nested dissection* [34]. For the remainder of this paper, we shall simply use nested dissection to refer to generalized nested dissection.

A separator is a global property of the graph. The quality of a nested dissection ordering depends crucially on the quality of its separators. With good separators, nested dissection gives high quality orderings for a large class of graphs. However, on a grid graph with a large aspect ratio, e.g., an $h \times k$ grid where h is appreciably smaller than k, a nested dissection ordering is inferior to a minimum degree ordering. A good separator is not enough; the ordering of the vertices found in the different levels of the separators is important. In many cases, the ordering given by the recursive ordering of the separators in nested dissection can be improved.

2.2. Shortcomings of minimum degree and nested dissection. To illustrate the inconsistent ordering quality from the minimum degree (MMD) [35] and the

[1] The elimination tree is a useful tool in the study of sparse matrix factorization. See [38] for a survey article.

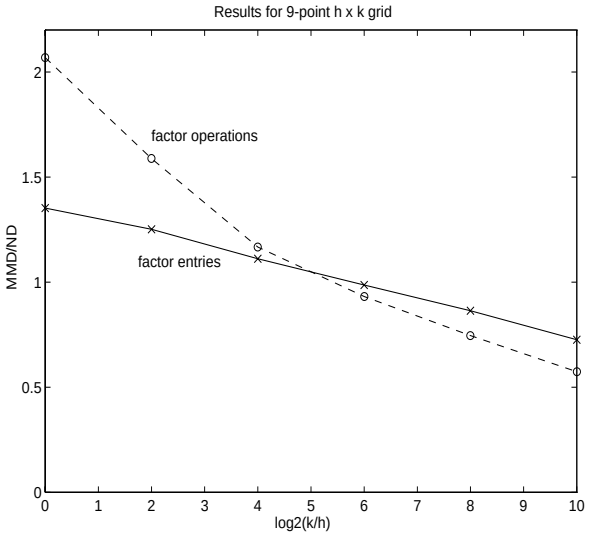


FIG. 1. *Comparison of MMD and ND for 2-D regular grids of different aspect ratios.*

grid-based nested dissection (ND) [15] algorithms, we apply the two algorithms on a sequence of rectangular grids of increasing aspect ratios. We run the orderings on grids of the following sizes: 128×128, 64×256, 32×512, 16×1024, 8×2048, and 4×4096, respectively, so that the number of unknowns are the same ($2^{14} = 16384$).

In Figure 1, we plot the performance ratio of MMD/ND versus the base-2 logarithm of the aspect ratios of the rectangular grids. The variation in performance is rather drastic. For the grid of unit aspect ratio, ND outperforms MMD by a factor of two in factorization operations. On the other end, for grids of large aspect ratios, MMD is better than ND in operations by a factor of close to two.

Such performance variations can also be found in practical sparse matrix problems. In section 5 we will compare minimum degree, nested dissection, and multisection on a set of matrices from the Harwell–Boeing sparse matrix collection [12]. For seven of the 16 matrices minimum degree generates an ordering with fewer operations than our nested dissection ordering. Part of this can be explained by the aspect ratio of the graphs of the matrices.[2] For example, one of the test matrices, BCSSTK25, is a finite element model of a tall building and has a graph with a large aspect ratio.

The shortcomings of the minimum degree ordering are largely due to the local nature of the algorithm. Selecting a vertex to eliminate, based on the local degree information, can often lead to less-than-desirable choices. Berman and Schnitger [8] have shown that there is a minimum degree sequence for the square grid so that the resulting ordering has factor nonzeros and operation counts an order of magnitude more than the optimal. By construction, their less-than-optimal ordering generates separators with a severe "fractal" nature. Virtually any minimum degree ordering has this property although to a lesser extent.

[2]The aspect ratio for a geometric object can be loosely defined using a major axis and cross-sections perpendicular to the axis. Both concepts can be extended to general graphs, where the Euclidean distance metric is replaced by the natural distance metric of a graph, namely, the distance between two vertices is the length of the shorted path connecting them.

On the other hand, the shortcoming of nested dissection can be best explained by its performance on problems of large aspect ratios. The experimental results in Figure 1 on rectangular grids of varying aspect ratios show that the difference in performance is quite significant. Since we are using the best separator (best in terms of both the separator size and the component balance) on the grid at each step of nested dissection, we cannot attribute the problem to the quality of the separators. The problem is with the way the last few levels of separators are numbered. Indeed, our approach of using multisectors provides a more effective way of numbering the vertices associated with these separators.

3. The multisection ordering algorithm. In nested dissection, a separator in the form of a *bisector* is used to split the given graph into two subgraphs where each subgraph is ordered recursively. The vertices associated with the bisector usually form a clique in the filled graph. If the separator is minimal and each of the two subgraphs is connected, then the bisector forms a clique in the filled graph. Our approach uses the notion of a *multisector* separator, which splits the given graph into a number of subgraphs. In general, the multisector vertices induce a sparse submatrix in the filled graph. Within this framework, we can, therefore, view nested dissection as a *multilevel bisection* scheme. On the other hand, this new approach is a *bilevel multisection* scheme. For simplicity, we shall simply refer to it as *multisection*.

3.1. Domain decomposition and the multisector. Multisection is closely related to the notion of domain decomposition, which we now formally define. Let A be an irreducible matrix with symmetric structure and let $G = (V, E)$ be the undirected graph of the structure of A. V is the set of vertices and $E \subseteq V \times V$ is the set of edges. Edge (i, j) is in E if and only if $a_{i,j} \neq 0$. For a subset $U \subseteq V$, the *boundary* of U, written $Adj(U)$, is the set of all vertices adjacent to those in U but not including any in U, i.e., $Adj(U) = \{v \notin U \mid (u, v) \in E \text{ for some } u \in U\}$.

Consider a partition of the vertex set V:

$$V = \Phi \cup \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_M,$$

where each Ω_i is a *domain* and Φ is the set of *interface* vertices. Each domain Ω_i is a connected subgraph of G whose boundary $Adj(\Omega_i)$ is contained in Φ. This partition of V induces a block partition of the matrix A,

$$A = \begin{bmatrix} A_{\Omega_1,\Omega_1} & & & A_{\Omega_1,\Phi} \\ & \ddots & & \vdots \\ & & A_{\Omega_M,\Omega_M} & A_{\Omega_M,\Phi} \\ A_{\Phi,\Omega_1} & \cdots & A_{\Phi,\Omega_M} & A_{\Phi,\Phi} \end{bmatrix}.$$

Since the domains are separated from one another by the set Φ, we shall also refer to Φ as a *multisector*, for it generalizes the notion of a bisector. Without loss of generality, we shall assume that the multisector partition is nontrivial; that is, $M > 1$ and Φ is nonempty.

Given this domain decomposition, we impose one condition on the ordering: *All vertices in the domains are numbered before any vertex in the multisector.* Since the domains are isolated from each other, each can be ordered independently.

We use the constrained minimum degree algorithm [37] to order the vertices in each domain. For domain Ω_i, we construct the graph $G_i = (\Omega_i \cup Adj(\Omega_i), E_i)$ where

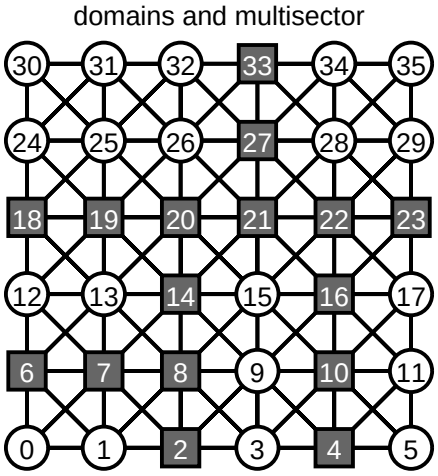


FIG. 2. *A multisector* $\Phi = \{2, 4, 6, 7, 8, 10, 14, 16, 18, 19, 20, 21, 22, 23, 27, 33\}$.

$E_i = E \cap (\Omega_i \times (\Omega_i \cup Adj(\Omega_i)))$ contains all edges $(u, v) \in E$ where $u \in \Omega_i$. While the vertices in $Adj(\Omega_i)$ contribute to the degrees of the vertices in Ω_i, only vertices in Ω_i are allowed to be eliminated.

To order the multisector vertices we consider $G_{V\setminus\Phi}$, the *elimination graph* of G after all vertices in the domains have been eliminated. Note that $G_{V\setminus\Phi}$ is also the graph of the *Schur complement matrix*

$$A_{\Phi,\Phi} - \sum_{i=1}^{M} A_{\Phi,\Omega_i} A_{\Omega_i,\Omega_i}^{-1} A_{\Omega_i,\Phi}.$$

From this equation it is clear that the elimination graph $G_{V\setminus\Phi}$ (the structure of the Schur complement matrix) does not depend on the ordering of domain vertices. Therefore, the ordering of the multisector vertices in this elimination graph can proceed independently of the ordering of the domain vertices.

Figure 2 contains an example of a domain decomposition of a 6×6 grid graph. The vertices are partitioned into six domains and a multisector:

$$\Phi = \{2, 4, 6, 7, 8, 10, 14, 16, 18, 19, 20, 21, 22, 23, 27, 33\}.$$

Multisector vertices are represented by squares and domain vertices by circles.

3.2. Compressed graphs. Often the time required to find an ordering can be significantly reduced by taking advantage of *indistinguishable vertices*. For a given graph, two vertices are said to be *indistinguishable* if they are adjacent and have exactly the same set of neighbors (including themselves). The elimination graph $G_{V\setminus\Phi}$ of a domain decomposition usually has many fewer indistinguishable vertices than original vertices. This is important since the execution time of the minimum degree ordering depends on the number of indistinguishable vertices instead of the number of original vertices. Figure 3 contains the Schur complement matrix associated with the domain decomposition of Figure 2. There are 16 vertices and 10 indistinguishable ones. For example, $\{4, 10, 16\}$ forms a set of indistinguishable vertices, since they all

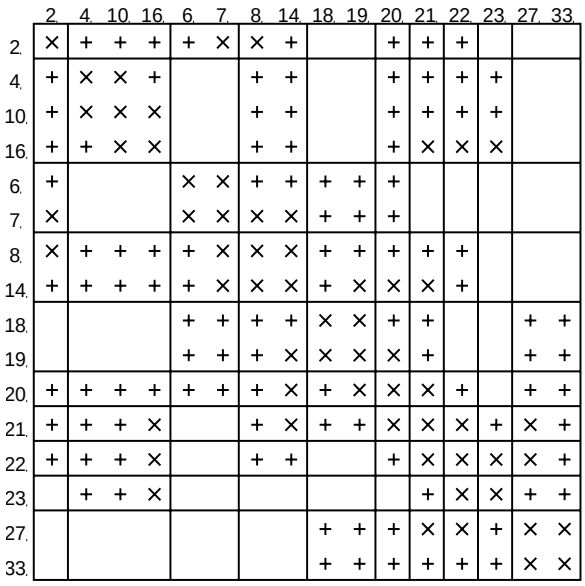


FIG. 3. *The Schur complement matrix has* 16 *vertices and* 10 *indistinguishable vertices. Original nonzero entries are denoted by '×,' fill entries by '+.'*

have the same adjacent set $\{2, 8, 14, 20, 21, 22, 23\}$ in the elimination graph. Indistinguishability is an equivalence relation defined on the original vertices, and so induces a partition of V.

Let $\mathbf{V}$ be any general partition of the original vertices V. The weight of $\mathbf{v} \in \mathbf{V}$, written $w(\mathbf{v})$, is the number of original vertices contained in $\mathbf{v}$. Similarly, the weight of the pair $(\mathbf{u}, \mathbf{v})$, written $w(\mathbf{u}, \mathbf{v})$, is the number of distinct edges $(u, v) \in E$ where $u \in \mathbf{u}$ and $v \in \mathbf{v}$. For a given partition $\mathbf{V}$ of V, the *compressed graph* induced by the partition is the graph $(\mathbf{V}, \mathbf{E})$, where $\mathbf{E}$ is the set of pairs $(\mathbf{u}, \mathbf{v}) \subseteq \mathbf{V} \times \mathbf{V}$ with nonzero weights.

In particular, the indistinguishability relation induces a partition $\mathbf{V}$ of the vertices, which in turn defines a special block partition of the original matrix A. To each edge $(\mathbf{u}, \mathbf{v})$ is associated a submatrix $A_{\mathbf{u},\mathbf{v}}$ of the original matrix, namely, the submatrix whose rows correspond to vertices in $\mathbf{u}$ and whose columns correspond to vertices in $\mathbf{v}$. The weight of the edge $(\mathbf{u}, \mathbf{v})$ is the number of nonzero entries in the corresponding submatrix. There is one important relation between the weights of an edge and its two incident vertices in the compressed graph, namely, $w(\mathbf{u}, \mathbf{v}) = w(\mathbf{u}) \cdot w(\mathbf{v})$, i.e., each submatrix $A_{\mathbf{u},\mathbf{v}}$ induced by the block partition is either dense or zero. Furthermore, there is no partition of the vertices with smaller cardinality for which $w(\mathbf{u}, \mathbf{v}) = w(\mathbf{u}) \cdot w(\mathbf{v})$ holds for all edges $(\mathbf{u}, \mathbf{v}) \in \mathbf{E}$.

In our software, instead of ordering the graph $G = (V, E)$ of the matrix A, we order $\mathbf{G} = (\mathbf{V}, \mathbf{E})$, the *natural compressed graph* [1], [11], of A, where $\mathbf{V}$ is the partition induced by the indistinguishability relation. Table 2 in section 5 contains some statistics on the sizes of the unit weight graph $G = (V, E)$ and the weighted graph $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ for some structural analysis matrices from the Harwell–Boeing test collection. The ratio $|\mathbf{E}|/|E|$ can be as high as 10-30 for these matrices, and since the complexity of the ordering process contains an $O(|E|)$ or $O(|\mathbf{E}|)$ term, the ordering times for the compressed graph can be appreciably smaller than those for the original graph.

3.3. A family of multisection ordering algorithms. Following is a skeleton of the multisection ordering scheme based on the notion of a domain decomposition.

```
MS (ordering method-1, ordering method-2):
    Given a domain decomposition (Φ, Ω_1, Ω_2, ... Ω_M) of V;
    for each domain Ω_i do
        Order the graph G_i = (Ω_i, E_i) by ordering method-1;
    Form the elimination graph G_{V\Φ} and order by ordering method-2;
```

A multisection (MS) ordering is defined by three choices:

1. How to determine the domain decomposition?
2. What fill-reducing ordering to order the domains Ω_i?
3. What fill-reducing ordering to order the multisector?

In the literature, there are a number of existing ordering schemes using this multisection approach.

• One-way nested dissection–MS(PROFILE, PROFILE). The *one-way dissection* scheme by George [16] chooses a set of equally spaced parallel lines as its multisector of a regular grid graph. Each component in the remaining graph is ordered by a profile ordering. The elimination graph associated with the multisector is also numbered by a profile ordering. We can, therefore, view one-way dissection as a MS(PROFILE, PROFILE) ordering scheme. George provided experimental and theoretical results to show that one-way dissection can be better than nested dissection for grid graphs with large aspect ratios. In his master thesis [40], Ng has considered the recursive use of the one-way dissection approach. This can also be viewed as using some form of multisector.

• Local nested dissection–MS(ND, PROFILE). The *local nested dissection* (LND) scheme in [9] carries the one-way dissection idea further. A rectangular grid of large aspect ratio is subdivided into roughly square domains by a set of parallel horizontal and vertical lines. Each square domain is ordered by nested dissection. The multisector defined by this set of lines is then numbered by a profile ordering. The LND scheme is, therefore, a MS(ND, PROFILE) ordering method.

• Incomplete nested dissection–MS(CMD, ND) and MS(MMD, ND). There are two generic forms of an *incomplete nested dissection* ordering. In both forms, the multisector is constructed using the recursive bisection process of nested dissection and the ordering of the multisector vertices follows the given nested dissection ordering. The difference lies in how the vertices in the domains are ordered, using either multiple minimum degree (MMD) [35] or constrained minimum degree (CMD) [37]. The latter algorithm usually generates a better ordering than the former on the domain subgraphs. There are many examples of incomplete nested dissection in the literature [3], [6], [7], [10], [19], [20], [22], [23], [28], [30], [33], [36], [42], [43], including three excellent state-of-the-art software packages, CHACO from Sandia National Laboratories [24], METIS from the University of Minnesota [29], and WGPP from IBM [21].

The above methods are all members of the multisection family of ordering algorithms. In the following sections we will compare incomplete nested dissection algorithms with a new method, MS(CMD, MMD), where vertices in the domains are ordered with constrained minimum degree and the Schur complement graph is ordered with MMD. We will refer to the MS(CMD, MMD) algorithm as *multisection*, in contrast with an incomplete nested dissection method (ND).

What remains is to specify how the multisector is created. In [6], the authors looked at two possibilities. The first method is to order the graph using MMD and use the elimination tree to extract a multisector. (Each domain is a subtree of the elimination tree; the multisector consists of all remaining vertices.) This multisector is then *smoothed* to remove the fractal nature of the separators that form the multisector. The second method, as described in [3], [6] performs recursive bisection on the graph until the subgraphs are a certain size, then take the multisector to be the union of the separators.

Subsequent to [6], Rothberg independently discovered the MS method using the second technique [45]. In his work, a multisector was formed of the separators obtained from the CHACO code [24] and automatic ND from SPARSPAK [17]. Multisection consistently performed as well or better than the ND algorithm that generated the multisector, evidence that supports our results in section 5 where we obtain a multisector from the METIS software package [29] and our own ND software [4].

4. Experimental results on regular grids. In this section we present some experimental results for the MS(CMD, MMD) ordering algorithm on regular grids. For grid problems it is easy to construct an ideal set of multisectors. In this way, we can study the effectiveness of MS(CMD, MMD) ordering when compared with some theoretically-optimal orderings.

4.1. Square and cubic regular grids. We first consider 9-point operators on two-dimensional $k \times k$ grids and 27-point operators on three-dimensional $k \times k \times k$ grids. For these graphs, the ND ordering [15] gives the best results (aside from some minor variations on very small grids due to edge effects). For a square $k \times k$ grid, ND first bisects the grid with a vertical separator creating two subgrids, each approximately $k/2 \times k$ in size. Each of these subgrids is bisected by a horizontal separator forming a total of four $k/2 \times k/2$ subgrids. The process repeats on each of the subgrids recursively. The separators are vertical or horizontal lines of grid points.

We construct a multisector in a similar way, composed of horizontal and vertical grid lines (planes in 3-D) that span the grid. The simplest multisector we call Φ_2, which splits each side of the grid in two with a single separator in each grid direction. There are four domains in 2-D, each approximately $k/2 \times k/2$ in size. For a 3-D grid there are eight domains, each approximately $k/2 \times k/2 \times k/2$ in size. The multisector Φ_3 splits the 2-D grid into nine domains, each approximately $k/3 \times k/3$ in size. In general, the Φ_m multisector creates $M = m^2$ domains in 2-D, ($M = m^3$ domains in 3-D), each domain is roughly k/m along each side. Note that the multisector Φ_1 is empty and the entire grid is one domain. We can parameterize the MS(CMD, MMD) ordering by the Φ_m multisector and Φ_1 corresponds to MMD on the original grid graph.

Our first experiment is to fix the number of domains and let the grid sizes grow. For 2-D $k \times k$ grids we looked at $1 \leq m \leq 7$. For 3-D $k \times k \times k$ grids we looked at $1 \leq m \leq 4$. Figure 4 contains four plots, results for 2-D grids at the top and 3-D grids on the bottom. The ratio of MS(CMD, MMD) factor entries to those of ND are found on the left and factorization operations on the right.

For ND, the number of factor entries is $O(k^2 \log k)$ in 2-D and $O(k^4)$ in 3-D; the number of factorization operations is $O(k^3)$ in 2-D and $O(k^6)$ in 3-D. We have scaled the number of factor entries and operations appropriately. In each plot the bottom curve is ND while the top curve is MMD. The MS curves are found between ND and MMD. As k grows, the relative performance of MMD versus ND becomes appreciably worse, more for factorization operations than factor entries and more for 3-D than 2-D.

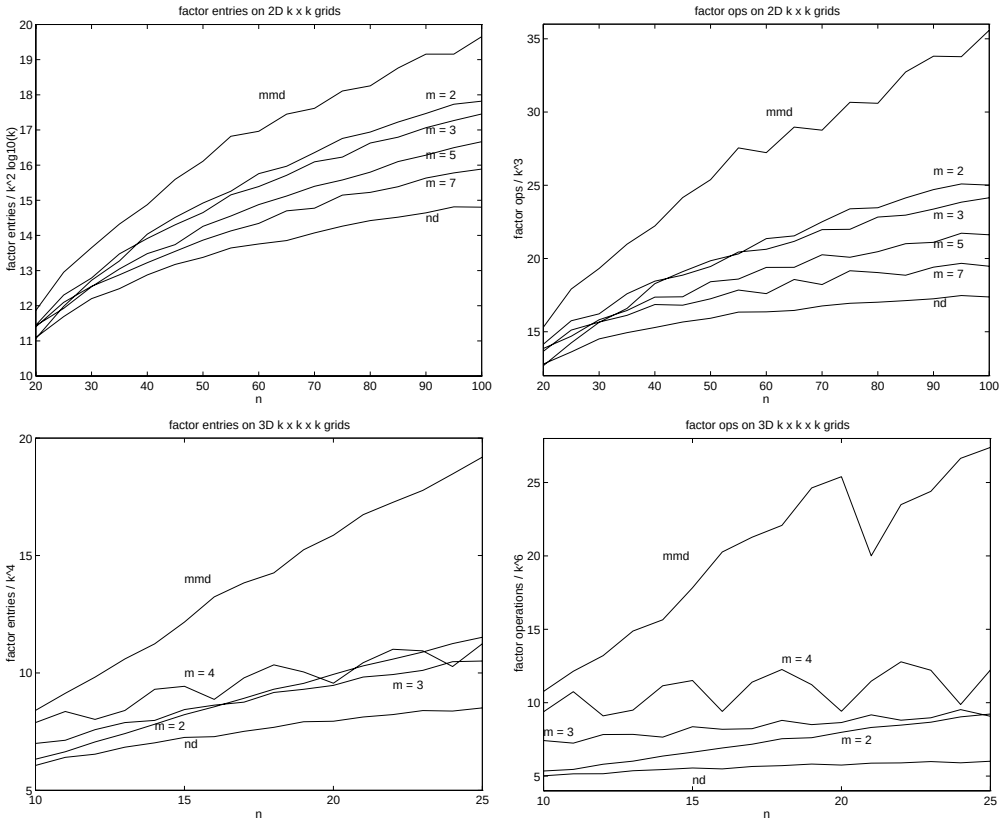


FIG. 4. *Multisection vs. nested dissection on square and cubic regular grids.*

The difference between MS and ND grows at a much smaller rate. For 2-D grids there is a steady improvement as m increases. For 3-D grids the smaller values of m are better; no doubt this is due to the relatively larger portion of factor entries and operations attached to the top level separators.

4.2. Rectangular quadrilaterals and hexahedra. Multisection performs fairly well when compared to ND on square and cubic regular grids. We now turn to grids with large aspect ratios and compare MS to local nested dissection (LND) [9], the best ordering for rectangular grids.

Table 1 presents some results for a 255×31 2-D grid and a $127 \times 15 \times 15$ 3-D grid. Both grids have an aspect ratio of 8, large enough to make LND clearly better than ND, but not large enough to make minimum degree better than nested dissection.

The results of MMD, ND, and MS are given relative to LND. The performance of ND relative to LND is virtually the same in two and three dimensions; ND requires around 10 percent more factor entries and 40 percent more factorization operations. However, MMD shows its strong dependence on the dimensionality of the graph for its 3-D performance is much worse than that for 2-D.

What is important to note is that MS will generate good orderings with many different multisectors. We have observed this behavior across a wide range of matrices; the quality of the MS ordering is not strongly dependent on the number

TABLE 1

Comparing multiple minimum degree (MMD), nested dissection (ND), multisection (MS), against local nested dissection (LND). The 255×31 *grid is partitioned by an* $m_1 \times m_2$ *grid of roughly square domains. The* $127 \times 15 \times 15$ *grid is partitioned by an* $m_1 \times m_2 \times m_3$ *grid of roughly cubic domains. All statistics are relative to LND.*

255×31 grid				
METHOD			NZF	OPS
MMD/LND			1.24	1.73
ND/LND			1.11	1.41
MS/LND				
m_1	m_2	M		
16	2	32	1.14	1.35
24	3	72	1.08	1.22
32	4	128	1.08	1.20
40	5	200	1.08	1.23
48	6	288	1.08	1.25

$127 \times 15 \times 15$ grid					
METHOD				NZF	OPS
MMD/LND				1.92	3.86
ND/LND				1.12	1.42
MS/LND					
m_1	m_2	m_3	M		
8	1	1	8	1.78	3.44
16	2	2	64	1.12	1.15
24	3	3	144	1.23	1.67

of domains induced by the multisector. Furthermore, additional experiments have also shown that the ordering quality is also relatively insensitive to the shape of domains.

5. Experimental results on some sparse matrices from structural analysis. The experimental results in section 4 on the 2D and 3D-grid problems suggest that the multisection ordering MS(CMD, MMD) can lead to very competitive orderings. The multisectors used are based on the geometry of the grids, and so can be regarded as the best we can get for a specified number of domains. For general sparse matrix problems, the MS ordering algorithm depends on the use of an appropriate domain decomposition. To avoid the "fractal" nature of the separators from MMD, each domain should have a "smooth" boundary.

5.1. Finding a domain decomposition via recursive bisection. A simple domain decomposition method can be formulated based on incomplete nested dissection. The vertex set V of the initial graph is decomposed into two or more connected subgraphs by removing a bisector S. The connected components of $V \setminus S$ are recursively bisected until each remaining subgraph is smaller than some prescribed size. Each remaining subgraph is a domain in the domain decomposition. The quality of the resulting domain decomposition depends on the method to find bisectors in the recursive steps.

Recently, there have been a number of published papers and software codes that find a partition of a given graph. Notable examples include the CHACO [24], METIS [29], and WGPP [21] software packages that use a multilevel approach to find a graph bisector.

We have developed a software code called DDSEP [4] that partitions a graph in three steps.

1. *Find an initial multisector.* Construct a domain decomposition of the graph by "growing" domains from random seed vertices.
2. *Find an initial separator formed of multisector vertices.* Form the *domain/segment* graph [4] where each segment is a subset of the multisector vertices. Apply a block version of the Kernighan–Lin scheme [32] on the domain/segment graph to obtain a graph bisector composed of segments.
3. *Improve the bisector.* Improve the bisector using graph matching

TABLE 2
Statistics for Harwell–Boeing matrices.

	ORIGINAL		COMPRESSED		MMD	
MATRIX	$\|V\|$	$\|E\|$	$\|\mathbf{V}\|$	$\|\mathbf{E}\|$	NZF/10^3	OPS/10^6
BCSSTK15	3948	113868	3948	113868	663	172
BCSSTK16	4884	285494	1778	36502	742	146
BCSSTK17	10974	417676	5219	81062	1141	201
BCSSTK18	11948	137142	10926	122177	657	138
BCSSTK23	3134	42044	2930	35256	461	142
BCSSTK24	3562	156348	892	12756	296	38
BCSSTK25	15439	236802	13183	161964	1544	339
BCSSTK29	13992	605496	10202	313846	1721	424
BCSSTK30	28924	2014568	9289	222884	3731	869
BCSSTK31	35588	1145828	17403	288806	5160	2411
BCSSTK32	44609	1970092	14821	226974	5175	1048
BCSSTK33	8738	583166	4344	164284	2656	1300
BCSSTK35	30237	1419926	6611	65934	2782	406
BCSSTK36	23052	1120088	4351	37166	2766	618
BCSSTK37	25503	1115474	7093	88924	2831	558
BCSSTK39	46772	2042522	10140	81762	7671	2194

techniques [36] or the Dulmage–Mendelsohn decomposition [13], [41] if the graph has unit weight vertices, or by using a generalized Dulmage–Mendelsohn decomposition [5] or max flow solver [5], [25] if the graph is weighted.

The DDSEP software has been demonstrated to be quite effective in finding good partitions of general connected graphs and also efficient in terms of execution time [4]. We now compare the domain decomposition approach DDSEP and the multilevel approach in METIS to find ND and MS orderings.

5.2. Results on practical structural problems. We have selected a set of practical sparse matrices arising from structural analysis problems from the Harwell–Boeing collection [12]. Table 2 provides a list of the problems and their characteristics. The column labeled ORIGINAL contains the number of vertices $|V|$ and the number of edges $|E|$ of the original given graph. We have also applied the graph compression technique in [1] to identify the indistinguishable vertices in the original graph. The column labeled COMPRESSED gives the number of vertices $|\mathbf{V}|$ and the number of edges $|\mathbf{E}|$ of the compressed graph. All of our ordering software works with the compressed graph, and this often results in significantly decreased ordering times when compared with using the original graph.

The last two columns in Table 2 present the number of factor entries and operations (both additions and multiplications) when the matrices are ordered using our MMD software, scaled by 10^3 and 10^6, respectively. For each matrix we made 21 runs of MMD where each run began with a randomly, symmetrically permuted matrix. Table 2 contains the median values of these runs. When we compare the ordering quality of the ND and MS methods, we scale their statistics by the corresponding MMD values.

Table 3 contains statistics for two ND algorithms—one from METIS and one using our DDSEP software—and MS where the multisector has been obtained from either METIS or DDSEP. Again, we made 21 runs for each algorithm and matrix and present the median value, scaled by the MMD factor entries (NZF) and factorization operations (OPS). An entry in the table that is greater (less) than one means that the

TABLE 3
A comparison of ND (nested dissection) and MS (multisection) relative to MMD (multiple minimum degree) using DDSEP and METIS to find the multisector via recursive bisection.

	NZF				OPS			
	ND		MS		ND		MS	
MATRIX	METIS	DDSEP	METIS	DDSEP	METIS	DDSEP	METIS	DDSEP
BCSSTK15	0.80	0.75	0.83	0.76	0.60	0.53	0.64	0.56
BCSSTK16	1.01	0.97	0.89	0.89	1.01	0.96	0.77	0.77
BCSSTK17	1.07	0.95	0.93	0.86	1.12	0.90	0.80	0.67
BCSSTK18	1.04	0.93	0.91	0.89	0.84	0.77	0.67	0.70
BCSSTK23	1.01	0.84	0.95	0.83	0.88	0.67	0.81	0.66
BCSSTK24	1.18	1.04	1.08	1.00	1.32	1.05	1.13	0.97
BCSSTK25	1.16	1.02	0.96	0.90	1.34	1.14	0.86	0.80
BCSSTK29	1.15	0.96	0.98	0.97	1.15	0.85	0.86	0.89
BCSSTK30	1.29	1.19	1.08	1.05	1.62	1.51	1.16	1.07
BCSSTK31	1.02	0.88	0.90	0.94	0.72	0.55	0.64	0.72
BCSSTK32	1.32	1.08	1.11	0.95	2.00	1.37	1.38	0.87
BCSSTK33	0.93	0.80	0.84	0.78	0.82	0.59	0.68	0.57
BCSSTK35	1.38	1.10	1.14	1.00	2.16	1.33	1.48	0.98
BCSSTK36	1.27	1.07	1.11	0.93	1.68	1.25	1.29	0.82
BCSSTK37	1.35	1.05	1.14	0.92	2.04	1.25	1.45	0.78
BCSSTK39	1.11	0.93	1.02	0.90	1.38	0.94	1.06	0.78
MEAN	1.13	0.97	0.99	0.92	1.29	0.98	0.98	0.79

ordering algorithm performed worse (better) than MMD. Both METIS[3] and DDSEP recursively split a subgraph until it has 100 or fewer vertices, (for a weighted graph, until the vertices in the subgraph have total weight less than 100).

Of the two ND algorithms, DDSEP consistently outperforms METIS. We believe that DDSEP produces better separators; see [4] for a comparison. There are three possible reasons:

- DDSEP is more flexible in enforcing a balance constraint than METIS; the latter tries to balance the size of the two subgraphs at the expense of a possible larger bisector. See [46] for a more complete discussion where evidence shows that it is effective to allow some imbalance in the partition to reduce the bisector size.
- DDSEP works exclusively with vertex bisectors while METIS finds an edge bisector and then extracts a vertex bisector; see [19], [20], [26] for a discussion of the drawbacks to finding a vertex separator from an edge separator.
- DDSEP uses a powerful algorithm (a generalized Dulmage–Mendelsohn decomposition or solving a max flow problem [5]) to smooth a bisector.

Both MS algorithms, the first using the multisector from METIS, the second using the multisector from DDSEP, consistently outperform their corresponding ND algorithms. Where ND is better than MS, the difference is not large. While there are seven matrices where ND using DDSEP does not produce as good an ordering as MMD, there is only one such case for MS using DDSEP.

Table 4 contains the ordering times for four out of the five methods. All ordering codes are written in C and were run on a Sparc20 using the `gcc` compiler with the `-O4` option. In general, METIS takes modest amounts of CPU times, but it is

[3]The options we used for METIS were recommended to us by the author, George Karypis, namely, SHEM (sorted heavy edge), BGKLR (combination of boundary greedy and boundary Kernighan–Lin), and GGPKL (graph growing followed by boundary Kernighan–Lin).

TABLE 4
Execution time for the ordering algorithms.

	ordering CPU time in seconds				portion of factorization time			
		ND		MS		ND		MS
MATRIX	MMD	METIS	DDSEP	DDSEP	MMD	METIS	DDSEP	DDSEP
BCSSTK15	1.18	1.33	3.39	3.45	11.7%	13.2%	33.6%	34.2%
BCSSTK16	0.27	2.76	1.40	1.39	3.1%	32.1%	16.3%	16.2%
BCSSTK17	0.82	5.19	4.43	4.45	7.6%	44.0%	37.5%	37.7%
BCSSTK18	2.50	1.81	10.38	10.47	30.9%	22.4%	128.2%	129.3%
BCSSTK23	0.90	0.77	1.95	1.95	10.7%	9.2%	23.2%	23.2%
BCSSTK24	0.07	1.32	0.58	0.57	3.2%	60.0%	26.4%	25.9%
BCSSTK25	3.34	4.79	11.62	11.42	7.1%	10.2%	24.6%	24.2%
BCSSTK29	1.77	6.89	11.22	11.24	7.1%	27.7%	45.1%	45.1%
BCSSTK30	1.66	23.81	11.08	11.15	3.3%	46.6%	21.7%	21.8%
BCSSTK31	4.74	18.51	20.04	20.27	3.3%	13.1%	14.1%	14.3%
BCSSTK32	2.75	28.63	16.89	17.08	4.5%	46.5%	27.4%	27.7%
BCSSTK33	1.04	18.51	5.75	5.78	1.4%	24.2%	7.5%	7.6%
BCSSTK35	0.85	17.69	6.50	6.57	3.6%	74.0%	27.2%	27.5%
BCSSTK36	0.74	13.72	4.20	4.23	2.0%	37.7%	11.5%	11.6%
BCSSTK37	0.93	14.22	6.81	6.82	2.8%	43.4%	20.8%	20.8%
BCSSTK39	1.28	29.99	10.09	10.18	1.0%	23.2%	7.8%	7.8%

penalized because it cannot work with the compressed graph.[4] Compare the times for BCSSTK15 where the compressed graph is identical to the original graph. For this matrix, METIS is almost three times as fast as DDSEP. We have observed this general tendency across the entire set of test matrices; DDSEP is usually a factor of two or more slower than METIS on the original graph. Part of this difference is due to the more powerful smoother, but part is because DDSEP trades more computation for reduced working storage. DDSEP uses only $O(|\mathbf{V}|\log(|\mathbf{V}|))$ working storage and does not replicate or destroy the input structure of the graph.

The ordering times for ND and MS with DDSEP include the time to order the vertices in the domains using our MMD software. The time in MS to order the vertices in the multisector is relatively small, and so we see the ND and MS ordering times are almost identical (again the median of 21 runs). It is clear that the bulk of the ordering time is spent finding the multisector via nested dissection.

Table 4 also provides the ordering times as the percentage of time needed for the numerical factorization of the MMD ordering. Our multifrontal factorization code from [2] consistently achieves 15–20 mflops for this collection of matrices. The MMD ordering time is a small percentage of the factorization time for the larger problems, while the ND and MS times can be up to five times greater. For all but four matrices, the MS ordering time takes less than one third of the factorization time. All the matrices in Table 4 are small to moderate in size. For larger matrices that we see in practice, up to two million degrees of freedom, the MMD ordering time is a very small fraction of the factorization time. The cost of the ND and MS orderings is also negligible.

The quality of the ND and MS orderings is somewhat dependent on the depth to which the ND is taken. Table 5 presents some statistics for three of the matrices. We have varied the maximum domain size (domain weight for a compressed graph) that defines when a subgraph will be split. Doubling the maximum domain size roughly means reducing the number of levels in the separator tree by one. The results

[4]METIS first finds an edge separator and then extracts the vertex separator using the graph matching algorithm from [41] that does not take into account any vertex weights.

TABLE 5

The influence of maximum domain size. $|\Omega_{max}|$ is the upper bound on the weight of a domain. Factor entries (NZF) and operation counts (OPS) are relative to multiple minimum degree. CPU times are in seconds.

		NZF		OPS		CPU	
MATRIX	$\|\Omega_{\max}\|$	ND	MS	ND	MS	ND	MS
BCSSTK31	100	0.88	0.94	0.56	0.73	20.04	20.27
	200	0.88	0.90	0.57	0.62	17.38	17.49
	400	0.90	0.88	0.59	0.59	15.27	15.29
	800	0.90	0.88	0.61	0.60	13.48	13.62
	1600	0.91	0.88	0.65	0.61	12.08	12.16
BCSSTK37	100	1.05	0.92	1.25	0.78	6.81	6.82
	200	1.07	0.95	1.30	0.85	5.61	5.59
	400	1.08	0.98	1.32	0.94	4.82	4.82
	800	1.12	1.05	1.44	1.12	4.05	4.04
	1600	1.14	1.09	1.54	1.32	3.39	3.39
BCSSTK39	100	0.93	0.90	0.94	0.78	10.09	10.18
	200	0.93	0.89	0.93	0.77	7.88	7.92
	400	0.96	0.91	0.98	0.79	6.66	6.72
	800	0.99	0.92	1.02	0.81	5.56	5.57
	1600	1.01	0.94	1.09	0.85	4.66	4.66

for BCSSTK31 show that ND improves as one reduces the maximum domain size but the MS ordering becomes worse. For BCSSTK37 both orderings improve as the maximum domain size decreases, while for BCSSTK39 the ordering quality is more flat throughout the parameter range. Note that the bulk of the ordering times are spent finding the separators at the highest levels. In general, the ordering quality of MS tends to be less sensitive to the number of levels of separators that are used to construct the multisector than ND, although this is problem dependent.

6. Concluding remarks. In this paper, we have introduced *multisection*, a robust ordering method using the notion of multisectors. We have demonstrated that it produces consistently high quality orderings for graphs of different characteristics. Its performance compares favorably with the popular minimum degree ordering MMD, and a state-of-the-art generalized ND software METIS.

There are several directions for future work. Foremost is to find high quality multisectors at less cost than performing ND. In the past [6] we have found a multisector by first ordering the graph using MMD, extracting a multisector from the elimination tree, smoothing this multisector, and then ordering using MS, for a total cost of between two and three times a single MMD ordering. In general, the quality of these orderings lies somewhere between those of MMD and MS using ND to find the multisector. We have constructed decent multisectors using the *generalized pseudoextents* algorithm from [18]. Again, the ordering quality lies between that of MMD and MS using ND to find the multisector, but finding the multisector can take considerable time, particularly when the number of domains is large. Each of these two methods produce multisectors that are locally smooth, i.e., the boundary of each domain is smooth, but there is little or no *global* smoothness as is found in a multisector from ND. In other words, the multisector may not contain a good *global* bisector.

It appears that some type of global smoothness should be present, i.e., the multisector must contain smooth portions at a higher level than the boundary of a single domain. We feel that a viable approach is to form the multisector using recursive *multisection* of the graph, in the same spirit as the quadrisection and octasection

from [23] and the k-way partitioning from [31]. This has the potential to reduce the ordering time because fewer levels of recursion are required, and possibly improve the resulting ordering, for often a multisector with a small number of subgraphs has better properties than the equivalent multisector found by repeated application of bisection.

Table 5 illustrates a drawback to the MS(CMD, MMD) ordering, namely, that the ordering may be sensitive to the choice of multisector. One could evaluate the orderings for several multisectors chosen from a single domain/separator tree and choose the best. While this will amortize the time spent in the recursive bisection process, the portion of time in the ordering steps will soon dominate. However, one can evaluate a sequence of MS(ND, MMD) orderings (ND on the domains) on a sequence of *nested* multisectors in much less time than to evaluate each separately. We will report on this work in progress in a future paper.

Acknowledgments. We would like to thank Ed Rothberg and Bruce Hendrickson for helpful correspondence during the course of this research. Comments by the referees have improved the presentation of this paper.

REFERENCES

[1] C. Ashcraft, *Compressed graphs and the minimum degree algorithm*, SIAM J. Sci. Comput., 16 (1995), pp. 1404–1411.

[2] C. Ashcraft, R. G. Grimes, and J. G. Lewis, *Accurate symmetric indefinite linear equation solvers*, SIAM J. Matrix Anal. Appl., 19 (1998), to appear.

[3] C. Ashcraft and J. W. H. Liu, *A Partition Improvement Algorithm for Generalized Nested Dissection*, Tech. report BCSTECH-94-020, Boeing Computer Services, Seattle, WA, 1994.

[4] C. Ashcraft and J. W. H. Liu, *Using domain decompositions to find graph bisectors*, BIT, 37 (1997), pp. 506–534.

[5] C. Ashcraft and J. W. H. Liu, *Applications of the Dulmage–Mendelsohn decomposition and network flow to graph bisection improvement*, SIAM J. Matrix. Anal. Appl., 19 (1998), pp. 325–354.

[6] C. Ashcraft and J. W. H. Liu, *Generalized nested dissection: Some recent progress*, Mini Symposium 5th SIAM Conference on Applied Linear Algebra, Snowbird, UT, June 18, 1994.

[7] S. T. Barnard and H. D. Simon, *A fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems*, in Proc. 6th SIAM Conference on Parallel Processing for Scientific Computing, Norfolk, VA, 1993, pp. 711–718.

[8] P. Berman and G. Schnitger, *On the performance of the minimum degree ordering for Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 83–88.

[9] M. V. Bhat, W. G. Habashi, J. W. H. Liu, V. N. Nguyen, and M. F. Peeters, *A note on nested dissection for rectangular grids*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 253–258.

[10] T. Bui and C. Jones, *A heuristic for reducing fill-in in sparse matrix factorization*, in Proc. 6th SIAM Conference on Parallel Processing for Scientific Computing, 1993, pp. 445–452.

[11] A. C. Damhaug, *Sparse Solution of Finite Element Equations*, Ph.D. thesis, Norwegian Institute of Technology, 1992.

[12] I. S. Duff, R. G. Grimes, and J. G. Lewis, *Sparse matrix test problems*, ACM Trans. Math. Software, 15 (1989), pp. 1–14.

[13] A. L. Dulmage and N. S. Mendelsohn, *Coverings of bipartite graphs*, Canad. J. Math, 10 (1958), pp. 517–534.

[14] J. George and J. W. H. Liu, *The evolution of the minimum degree ordering algorithm*, SIAM Rev., 31 (1989), pp. 1–19.

[15] J. A. George, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.

[16] J. A. George, *An automatic one-way dissection algorithm for irregular finite element problems*, SIAM J. Numer. Anal., 17 (1980), pp. 740–751.

[17] J. A. George, J. W. H. Liu, and E. G. Ng, *User's guide for SPARSPAK: Waterloo Sparse Linear Equations Package*, Tech. report CS78-30 (revised), Department of Computer Sci-

ences, University of Waterloo, Waterloo, ON, Canada, 1980.
[18] T. Goehring and Y. Saad, *Heuristic Algorithms for Automatic Graph Partitioning*, Tech. report, Computer Science Department, University of Minnesota, Minneapolis, MN, 1995.
[19] A. Gupta, *Fast and Effective Algorithms for Graph Partitioning and Sparse Matrix Ordering*, Tech. report RC 20496 (90799), IBM T. J. Watson Research Center, Yorktown Heights, NY, 1996.
[20] A. Gupta, *Graph Partitioning Based Sparse Matrix Orderings for Interior Point Algorithms*, Tech. report RC 20467 (90480), IBM T. J. Watson Research Center, Yorktown Heights, NY, 1996.
[21] A. Gupta, *WGPP: Watson Graph Partitioning and Sparse Matrix Ordering Package: Users Manual*, Tech. report RC 20496 (90799), IBM T. J. Watson Research Center, Yorktown Heights, NY, 1996.
[22] M. T. Heath and P. Raghavan, *A Cartesian Nested Dissection Algorithm*, Tech. report UIUCDCS-R-92-1772, Department of Computer Science, University of Illinois, Urbana, IL, 1992.
[23] B. Hendrickson and R. Leland, *An Improved Spectral Graph Partitioning Algorithm for Mapping Parallel Computations*, Tech. report SAND92-1460, Sandia National Laboratories, Albuquerque, NM, 1992.
[24] B. Hendrickson and R. Leland, *The Chaco User's Guide*, Tech. report SAND93-2339, Sandia National Laboratories, Albuquerque, NM, 1993.
[25] B. Hendrickson and E. Rothberg, *Improving the runtime and quality of nested dissection ordering*, SIAM J. Sci. Comput., to appear.
[26] B. Hendrickson and E. Rothberg, *A Multi-Level Approach to Computing Node Separators for Nested Dissection Ordering*, Tech. report, Sandia National Laboratories, Albuquerque, NM, 1996, in preparation.
[27] A. J. Hoffman, M. S. Martin, and D. J. Rose, *Complexity bounds for regular finite difference and finite element grids*, SIAM J. Numer. Anal., 10 (1973), pp. 364–369.
[28] G. Karypis and V. Kumar, *A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs*, Tech. report TR 95-035, Department of Computer Science, University of Minnesota, Minneapolis, MN, 1995.
[29] G. Karypis and V. Kumar, *METIS: Unstructured Graph Partitioning and Sparse Matrix Ordering System*, Tech. report, Department of Computer Science, University of Minnesota, Minneapolis, MN, 1995.
[30] G. Karypis and V. Kumar, *A fast and high quality multilevel scheme for partitioning irregular graphs*, SIAM J. Sci. Comput., to appear.
[31] G. Karypis and V. Kumar, *Multilevel k-way Partitioning Scheme for Irregular Graphs*, Tech. report, Department of Computer Science, University of Minnesota, Minneapolis, MN, 1995.
[32] B. W. Kernighan and S. Lin, *An efficient heuristic procedure for partitioning graphs*, Bell System Tech. J., 49 (1970), pp. 291–307.
[33] C. E. Leiserson and J. G. Lewis, *Ordering for parallel sparse symmetric factorization*, in Parallel Processing for Scientific Computing, SIAM, Philadelphia, PA, 1989, pp. 27–31.
[34] R. J. Lipton, D. J. Rose, and R. E. Tarjan, *Generalized nested dissection*, SIAM J. Numer. Anal., 16 (1979), pp. 346–358.
[35] J. W. H. Liu, *Modification of the minimum degree algorithm by multiple elimination*, ACM Trans. Math. Software, 11 (1985), pp. 141–153.
[36] J. W. H. Liu, *A graph partitioning algorithm by node separators*, ACM Trans. Math. Software, 15 (1989), pp. 198–219.
[37] J. W. H. Liu, *On the minimum degree ordering with constraints*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1136–1145.
[38] J. W. H. Liu, *The role of elimination trees in sparse factorization*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 134–172.
[39] H. M. Markowitz, *The elimination form of the inverse and its application to linear programming*, Management Sci., 3 (1957), pp. 255–269.
[40] E. Ng, *On One-Way Dissection Schemes*, Master's thesis, Department of Computer Science, University of Waterloo, Waterloo, ON, Canada, 1979.
[41] A. Pothen and C. Fan, *Computing the block triangular form of a sparse matrix*, ACM Trans. Math. Software, 16 (1990), pp. 303–324.
[42] A. Pothen, H. Simon, and K. P. Liou, *Partitioning sparse matrices with eigenvectors of graphs*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 430–452.

[43] P. Raghavan, *Parallel Ordering Using Edge Contraction*, Tech. report CS-95-293, Department of Computer Science, University of Tennessee, Knoxville, TN, 1995.
[44] D. J. Rose, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Graph Theory and Computing, R. Read, ed., Academic Press, New York, 1972, pp. 183–217.
[45] E. Rothberg, *private communication*, 1995.
[46] E. Rothberg, *Exploring the tradeoff between imbalance and separator size in nested dissection ordering*. Silicon Graphics, 1996, submitted.
[47] W. F. Tinney and J. W. Walker, *Direct solutions of sparse network equations by optimally ordered triangular factorization*, J. Proc. IEEE, 55 (1967), pp. 1801–1809.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 3, pp. 833–845, July 1998

016

A PADÉ APPROXIMATION METHOD FOR SQUARE ROOTS OF SYMMETRIC POSITIVE DEFINITE MATRICES*

YA YAN LU†

Abstract. A numerical method for computing the square root of a symmetric positive definite matrix is developed in this paper. It is based on the Padé approximation of $\sqrt{1+x}$ in the prime fraction form. A precise analysis allows us to determine the minimum number of terms required in the Padé approximation for a given error tolerance. Theoretical studies and numerical experiments indicate that the method is more efficient than the standard method based on the spectral decomposition, unless the condition number is very large.

Key words. matrix square root, Padé approximation, prime fraction form

AMS subject classifications. 15A15, 41A21, 65F30

PII. S089547989731631X

1. Introduction.

The numerical computation of the square root of a matrix has been studied by a number of authors [12, 14, 4, 6, 7, 5, 11, 10, 13]. A popular approach is based on the Schur decomposition of the matrix [4, 7]. Iterative methods are also possible [12, 14, 6, 13]. For a symmetric positive definite matrix A, there is a unique symmetric positive definite square root (denoted by $\sqrt{A}$). For the given spectral decomposition

$$A = V\Lambda V^T, \tag{1.1}$$

where Λ is the diagonal matrix of the eigenvalues, V is the orthogonal matrix of the corresponding eigenvectors; the square root of A is given by

$$\sqrt{A} = V\sqrt{\Lambda}V^T. \tag{1.2}$$

Since $\sqrt{A}$ is also symmetric, it is only necessary to calculate its lower or upper triangular part. If the decomposition (1.1) is available, about n^3 additional operations are needed to calculate $\sqrt{A}$ by (1.2). For a full matrix A, the spectral decomposition (1.1) requires approximately $9n^3$ operations [8]. Therefore, the total number of arithmetic operations is around $10n^3$.

In this paper, we present a new method for computing $\sqrt{A}$ that requires about $10n^3/3+5n^2m/2$ operations, where m is an integer that depends on the desired relative accuracy ϵ and the spectral condition number κ of the matrix A. More precisely, we have

$$m \sim \frac{\kappa^{1/4}}{4}\sqrt{\ln\left(\frac{2}{\epsilon}\right)\ \ln\left(1+\frac{2}{\epsilon\sqrt{\kappa}}\right)}. \tag{1.3}$$

The new method should be more efficient than the standard procedure if the matrix A is not very ill conditioned. It turns out that parallel implementation of the new method is also very easy.

*Received by the editors February 10, 1997; accepted for publication (in revised form) by L. Reichel November 11, 1997; published electronically May 6, 1998. This research was partially supported by City University of Hong Kong research grant 9030458.

http://www.siam.org/journals/simax/19-3/31631.html.

†Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (mayylu@cityu.edu.hk).

Our method is based on the (m,m) diagonal Padé approximation [2] for the function $\sqrt{1+x}$ written in the prime fraction form [3, 9]

$$\sqrt{1+x} \approx 1 + \sum_{j=1}^{m} \frac{a_j^{(m)} x}{1 + b_j^{(m)} x}, \tag{1.4}$$

where

$$a_j^{(m)} = \frac{2}{2m+1} \sin^2 \frac{j\pi}{2m+1}, \quad b_j^{(m)} = \cos^2 \frac{j\pi}{2m+1}. \tag{1.5}$$

The basic steps of the new method are presented in the next section. For a given desired accuracy ϵ, a technique for choosing the integer m and a proper scaling parameter is developed in section 3. Numerical examples are presented in section 4.

2. The new method. For an $n \times n$ symmetric positive definite matrix A and a small positive parameter ϵ, our method for finding $\sqrt{A}$ is as follows: 1. Reduce the matrix A to a tridiagonal matrix by orthogonal similarity transformations. That is,

$$A = QTQ^T,$$

where Q is an orthogonal matrix and T is a symmetric tridiagonal matrix.

2. Find the largest and smallest eigenvalues of T, say $\lambda_1 > \lambda_n > 0$.
3. Determine the integer m and a scalar $\mu \in (\lambda_n, \lambda_1)$ as follows:
   (a) Find $t \in (0,1)$ from the following equation by Newton's method:

$$(\alpha\beta)^t + 1 = \tau(\alpha^t + \beta^t), \tag{2.1}$$

where

$$\alpha = \frac{2}{\epsilon} - 1, \quad \beta = 1 + \frac{2}{\epsilon\sqrt{\kappa}}, \quad \tau = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}, \quad \kappa = \frac{\lambda_1}{\lambda_n}. \tag{2.2}$$

   (b) Find m and μ by

$$m = \left\lceil \frac{1-t}{2t} \right\rceil, \quad \mu = \lambda_1 \left(\frac{\alpha^t - 1}{\alpha^t + 1} \right)^2. \tag{2.3}$$

4. For the symmetric tridiagonal matrix $X = T/\mu - I$, evaluate

$$R_m = \sqrt{\mu} \sum_{j=1}^{m} a_j^{(m)} (I + b_j^{(m)} X)^{-1} X,$$

where $a_j^{(m)}, b_j^{(m)}$ are the Padé coefficients listed in (1.5).

5. Evaluate the approximation of $\sqrt{A}$ by

$$\sqrt{A} \approx S_m = \sqrt{\mu} I + Q R_m Q^T. \tag{2.4}$$

For the matrix X defined above, we have $T = \mu(I + X)$. Therefore,

$$\sqrt{A} = Q\sqrt{T}Q^T = Q\sqrt{\mu(I+X)}Q^T = \sqrt{\mu} Q \sqrt{I+X} Q^T.$$

Based on the Padé approximation of $\sqrt{1+x}$, we have the following approximation for the square root of $I+X$:

$$\sqrt{I+X} \approx I + \sum_{j=1}^{m} a_j^{(m)} (I + b_j^{(m)} X)^{-1} X.$$

This gives rise to

$$\sqrt{A} \approx \sqrt{\mu} I + Q \left(\sqrt{\mu} \sum_{j=1}^{m} a_j^{(m)} (I + b_j^{(m)} X)^{-1} X \right) Q^T = \sqrt{\mu} I + Q R_m Q^T.$$

Step 1 above is exactly the same as in standard numerical methods for computing eigenvalues and eigenvectors of a symmetric matrix [8]. When the matrix A is full, Householder reflectors are usually used for this purpose. The matrix Q is not explicitly formed; only the related vectors for each Householder matrix are stored. The number of arithmetic operations required in this step is about $4n^3/3$. In our program, we use the LAPACK [1] routine `xSYTRD`.

Step 2 calculates the extreme eigenvalues of T (also of A). Many different methods can be used. The required number of operations should be $O(n)$. We use the bisection method `xSTEBZ` of LAPACK in our program. This step is necessary since we need the extreme eigenvalues to determine the optimal scaling parameter μ and the integer m in the Padé approximation (1.4). The time spent on this step is negligible.

The theory behind step 3 will be developed in the next section. A good initial guess for t (for solving (2.1) with Newton's method) is given in (3.4). The number of operations required in this step is $O(1)$. In section 3, it is proved that

$$||\sqrt{A} - S_m||_2 \leq \epsilon ||\sqrt{A}||_2 \tag{2.5}$$

for the chosen m and μ, where $||\cdot||_2$ denotes the matrix 2-norm.

Step 4 involves the sum of m matrices

$$X_j^{(m)} = (I + b_j^{(m)} X)^{-1} (\sqrt{\mu} a_j^{(m)} X).$$

We could initialize R_m as the zero matrix, then update R_m by

$$R_m := R_m + X_j^{(m)} \quad \text{for} \quad j = 1, 2, ..., m.$$

Only the lower (or upper) triangular part of the matrix needs to be calculated, since the matrices are symmetric. To find $X_j^{(m)}$, it is necessary to solve the system

$$(I + b_j^{(m)} X) X_j^{(m)} = \sqrt{\mu} a_j^{(m)} X. \tag{2.6}$$

The tridiagonal coefficient matrix $I + b_j^{(m)} X$ is symmetric positive definite, since T is positive definite, $0 < b_j^{(m)} < 1$, and

$$I + b_j^{(m)} X = (1 - b_j^{(m)}) I + \frac{b_j^{(m)}}{\mu} T.$$

The matrix $I + b_j^{(m)} X$ is always better conditioned than the original matrix A. Its spectral condition number can easily be found as

$$\text{cond}(I + b_j^{(m)} X) = \frac{\lambda_1 + \mu \tan^2(j\theta)}{\lambda_n + \mu \tan^2(j\theta)} \quad \text{for} \quad \theta = \frac{\pi}{2m+1}.$$

For large $\kappa = \lambda_1/\lambda_n$, the asymptotic formula (3.6) for μ is derived in section 3. Therefore,

$$\mathrm{cond}(I + b_j^{(m)} X) \sim \frac{D_j + \sqrt{\kappa}}{D_j + \sqrt{\kappa^{-1}}}, \quad \text{where} \quad D_j = \frac{\ln \alpha}{\ln \beta} \tan^2(j\theta)$$

for α, β given in (2.2). We conclude that for a typical j, the spectral condition number of the matrix $I + b_j^{(m)} X$ is $O(\sqrt{\kappa})$. For small j, the condition number is larger, but the coefficient $a_j^{(m)}$ is smaller. Therefore, the relatively large error in the numerical solution of $(I+b_j^{(m)} X)^{-1} X$ is reduced by the factor of $a_j^{(m)}$ in $X_j^{(m)}$. If the matrix A is not very ill conditioned, the integer m is not large (compared with n), and we expect that the matrices $X_j^{(m)}$ (for $j = 1, 2, ..., m$) and R_m can be accurately calculated.

To solve (2.6), we first calculate the decomposition $I + b_j^{(m)} X = LDL^T$, where L is a unit lower bidiagonal matrix and D is a diagonal matrix. This requires $O(n)$ operations. The columns of (2.6) can be solved afterwards. Since X is tridiagonal, the right-hand side of the kth column has at most three nonzero entries. Meanwhile, only the lower triangular part of the matrix $X_j^{(m)}$ is needed. These considerations lead to a reduction in the total number of operations required. The solution of column k requires about $4(n-k)$ operations. The summation of all k for $1 \le k \le n$ gives rise to the leading term $2n^2$. This is the total number of operations required to solve $X_j^{(m)}$. Since there are m such matrices and we have to add them together, the number of operations required to calculate R_m is thus $\frac{5}{2} m n^2$.

For a share memory multiprocessor computer, step 4 can be efficiently implemented. Clearly, the m matrices $X_j^{(m)}$ can be calculated independently on different processors. After that, the sum can be carried out recursively in pairs. When more processors are available, the columns of the matrix $X_j^{(m)}$ can also be computed concurrently once the LDL^T decomposition is completed. On a distributed memory multicomputer system, if the matrices $X_j^{(m)}$ are calculated in different processes, the communication cost to add them together may be too high. Alternatively, concurrency can be achieved by calculating a block of columns of R_m in each process.

When the matrix A is reduced to the tridiagonal matrix T in step 1, a sequence of Householder reflectors is used. We have $Q^T = H_{n-2} H_{n-3} \cdots H_2 H_1$, where H_k is an orthogonal matrix of the form $I - \tau_k v_j v_k^T$ for a scalar τ_k and some column vector v_k. In step 5, we calculate $Q R_m Q^T$ for a given symmetric matrix R_m. This can easily be achieved by applying the Householder reflectors in the reverse order. Using the same technique as in the reduction step, we have an efficient algorithm that requires only $2n^3$ operations to calculate $Q R_m Q^T$. Finally, the matrix S_m is obtained by adding the diagonals by $\sqrt{\mu}$.

The total number of arithmetic operations required by our method is $\frac{10}{3} n^3 + \frac{5}{2} n^2 m$. The main contributions come from steps 1, 4, and 5. The time spent on steps 2 and 3 are negligible.

3. Parameter selection. In this section, we study the selection of the integer m and the scaling parameter μ for a given desired accuracy ϵ and the given extreme eigenvalues λ_1, λ_n. This is closely related to the accuracy of the Padé approximation to the function $\sqrt{1+x}$. We first establish an exact formula for the error of this approximation.

THEOREM 3.1. *For any nonnegative integer m and $x > -1$, let*

$$E_m(x) = \sqrt{1+x} - 1 - \sum_{j=1}^{m} \frac{a_j^{(m)} x}{1 + b_j^{(m)} x}, \tag{3.1}$$

where $a_j^{(m)}$, $b_j^{(m)}$ are given as in (1.5). *Then*

$$E_m(x) = 2\sqrt{1+x}\frac{\gamma^{2m+1}(x)}{1+\gamma^{2m+1}(x)} \quad \textit{for} \quad \gamma(x) = \frac{\sqrt{1+x}-1}{\sqrt{1+x}+1}.$$

Proof. Let $f(x) = \sqrt{1+x} - 1$; we observe that

$$f(x) = \frac{x}{2 + f(x)}.$$

The continued fraction approximation to $f(x)$ can be introduced by

$$f_0(x) = 0, \qquad f_{k+1} = \frac{x}{2 + f_k(x)} \quad \text{for} \quad k = 0, 1, 2, ...$$

It is known [3] that the functions $\{f_2(x), f_4(x), f_6(x), ...\}$ are the diagonal Padé approximations of $f(x)$, and they can be written as

$$f_{2m}(x) = \sum_{j=1}^{m} \frac{a_j^{(m)} x}{1 + b_j^{(m)} x}.$$

The rational recursion for $\{f_k(x)\}$ can be solved. Let

$$R = \begin{bmatrix} 0 & x \\ 1 & 2 \end{bmatrix}$$

and the kth power of R be given by

$$R^k = \begin{bmatrix} r_{11}^{(k)} & r_{12}^{(k)} \\ r_{21}^{(k)} & r_{22}^{(k)} \end{bmatrix}.$$

Then it is easy to prove by induction that

$$f_k(x) = \frac{r_{11}^{(k)} f_0(x) + r_{12}^{(k)}}{r_{21}^{(k)} f_0(x) + r_{22}^{(k)}} = \frac{r_{12}^{(k)}}{r_{22}^{(k)}}.$$

The eigenvalues of R are

$$\sigma_1 = 1 + \sqrt{1+x}, \quad \sigma_2 = 1 - \sqrt{1+x}.$$

Writing down the corresponding eigenvectors, we obtain

$$R = \begin{bmatrix} x & x \\ \sigma_1 & \sigma_2 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} \begin{bmatrix} x & x \\ \sigma_1 & \sigma_2 \end{bmatrix}^{-1}.$$

Therefore,

$$R^k = -\frac{1}{2x\sqrt{1+x}} \begin{bmatrix} x(\sigma_1^k \sigma_2 - \sigma_1 \sigma_2^k) & x^2(\sigma_2^k - \sigma_1^k) \\ \sigma_1^{k+1}\sigma_2 - \sigma_1 \sigma_2^{k+1} & x(\sigma_2^{k+1} - \sigma_1^{k+1}) \end{bmatrix}.$$

This gives rise to

$$f_k(x) = \frac{r_{12}^{(k)}}{r_{22}^{(k)}} = \frac{x(\sigma_1^k - \sigma_2^k)}{\sigma_1^{k+1} - \sigma_2^{k+1}}.$$

Notice that $f(x) = x/\sigma_1$; we have

$$f_k(x) - f(x) = \frac{x[1-(\sigma_2/\sigma_1)^k]}{\sigma_1[1-(\sigma_2/\sigma_1)^{k+1}]} - \frac{x}{\sigma_1} = 2\sqrt{1+x}\frac{(\sigma_2/\sigma_1)^{k+1}}{1-(\sigma_2/\sigma_1)^{k+1}}.$$

For $k = 2m$ and $\sigma_2/\sigma_1 = -\gamma$, we get the desired result for $E_m(x) = f(x) - f_{2m}(x)$. □

If the Padé approximation for $\sqrt{1+x}$ is used to approximate $\sqrt{\lambda}$ for $\lambda \in [\lambda_n, \lambda_1]$, it is natural to scale λ by μ, write $\lambda = \mu(1+x)$, and then use the formula

$$\sqrt{\lambda} = \sqrt{\mu}\sqrt{1+x} \approx \sqrt{\mu}\left(1 + \sum_{j=1}^{m} \frac{a_j^{(m)} x}{1 + b_j^{(m)} x}\right) \quad \text{for} \quad x \in \left[\frac{\lambda_n}{\mu} - 1, \frac{\lambda_1}{\mu} - 1\right].$$

One immediate question asks: how to choose μ such that the maximum error for approximating $\sqrt{\lambda}$ on the interval $[\lambda_n, \lambda_1]$ is minimized. Furthermore, if a desired accuracy is given, what is the minimum m to obtain that accuracy? These questions are answered in the following theorem.

THEOREM 3.2. *For $\epsilon > 0$ and $\lambda_1 > \lambda_n > 0$, let m and μ be given by* (2.3)*; then*

$$\sqrt{\mu}\ |E_m(x)| \ \le \epsilon\sqrt{\lambda_1} \quad \textit{for} \quad \frac{\lambda_n}{\mu} - 1 \le x \le \frac{\lambda_1}{\mu} - 1.$$

Proof. For any integer $m \ge 0$ and $\mu \in (\lambda_n, \lambda_1)$, from Theorem 3.1, we have

$$E_m\left(\frac{\lambda_1}{\mu} - 1\right) = 2\sqrt{\frac{\lambda_1}{\mu}}\left[1 + \left(\frac{\sqrt{\lambda_1} + \sqrt{\mu}}{\sqrt{\lambda_1} - \sqrt{\mu}}\right)^{2m+1}\right]^{-1},$$

$$-E_m\left(\frac{\lambda_n}{\mu} - 1\right) = 2\sqrt{\frac{\lambda_n}{\mu}}\left[\left(\frac{\sqrt{\mu} + \sqrt{\lambda_n}}{\sqrt{\mu} - \sqrt{\lambda_n}}\right)^{2m+1} - 1\right]^{-1}.$$

The above formulas allow us to extend the definitions of $E_m(\lambda_1/\mu - 1)$ and $E_m(\lambda_n/\mu - 1)$ to arbitrary real number m. Now, for any fixed $m \ge 0$, we observe that there is a unique solution of μ in (λ_n, λ_1) such that

$$\sqrt{\mu}E_m\left(\frac{\lambda_1}{\mu} - 1\right) = -\sqrt{\mu}E_m\left(\frac{\lambda_n}{\mu} - 1\right). \tag{3.2}$$

This is so because $\sqrt{\mu}E_m(\lambda_1/\mu - 1)$ is a monotonically decreasing function of μ that takes a positive value at λ_n and is zero at λ_1, while $-\sqrt{\mu}E_m(\lambda_n/\mu - 1)$ is a monotonically increasing function of μ that is zero at λ_n and positive at λ_1. On the other hand, for any fixed μ in (λ_n, λ_1), both $E_m(\lambda_1/\mu - 1)$ and $E_m(\lambda_n/\mu - 1)$ are decreasing functions of m (converge to zero as $m \to \infty$). If we denote the value of both sides of (3.2) by ϵ_m, then ϵ_m is a decreasing function of m that converges to zero as $m \to \infty$.

Now, for a small $\epsilon > 0$, let m_* be the real number such that $\epsilon_{m_*} = \epsilon\sqrt{\lambda_1}$. We show that $t = 1/(2m_* + 1)$ satisfies (2.1). To simplify the notation, let $\kappa = \lambda_1/\lambda_n$ and $s = \sqrt{\mu}/\sqrt[4]{\lambda_1\lambda_n}$. This gives rise to

$$E_m\left(\frac{\lambda_1}{\mu} - 1\right) = 2\sqrt{\frac{\lambda_1}{\mu}}\left[1 + \left(\frac{\kappa^{1/4} + s}{\kappa^{1/4} - s}\right)^{2m+1}\right]^{-1},$$
$$-E_m\left(\frac{\lambda_n}{\mu} - 1\right) = 2\sqrt{\frac{\lambda_n}{\mu}}\left[\left(\frac{s + \kappa^{-1/4}}{s - \kappa^{-1/4}}\right)^{2m+1} - 1\right]^{-1}.$$

The conditions for m_* and μ are

$$E_{m_*}\left(\frac{\lambda_1}{\mu} - 1\right) = -E_{m_*}\left(\frac{\lambda_n}{\mu} - 1\right) = \epsilon\sqrt{\frac{\lambda_1}{\mu}}. \tag{3.3}$$

This leads to

$$\left(\frac{\kappa^{1/4} + s}{\kappa^{1/4} - s}\right)^{2m_*+1} = \frac{2}{\epsilon} - 1 = \alpha,$$
$$\left(\frac{s + \kappa^{-1/4}}{s - \kappa^{-1/4}}\right)^{2m_*+1} = 1 + \frac{2}{\epsilon\sqrt{\kappa}} = \beta,$$

or

$$\frac{\kappa^{1/4} + s}{\kappa^{1/4} - s} = \alpha^t, \quad \frac{s + \kappa^{-1/4}}{s - \kappa^{-1/4}} = \beta^t.$$

Solving s from the above two equations, we have

$$s = \frac{\alpha^t - 1}{\alpha^t + 1}\kappa^{1/4} = \frac{\beta^t + 1}{\beta^t - 1}\kappa^{-1/4}.$$

This gives rise to an equation for t:

$$(\sqrt{\kappa} - 1)(\alpha\beta)^t - (\sqrt{\kappa} + 1)(\alpha^t + \beta^t) + \sqrt{\kappa} - 1 = 0.$$

The above is the same as (2.1). From the earlier equation for s and $s = \sqrt{\mu}/\sqrt[4]{\lambda_1\lambda_n}$, we obtain

$$\mu = \lambda_1\left(\frac{\alpha^t - 1}{\alpha^t + 1}\right)^2.$$

It is straight forward to verify that (2.1) for t and the above formula for μ also imply the condition (3.3). Since both $E_m(\lambda_1/\mu - 1)$ and $E_m(\lambda_n/\mu - 1)$ are positive decreasing functions of m and $m \geq m_*$, we have

$$0 < E_m\left(\frac{\lambda_1}{\mu} - 1\right) \leq E_{m_*}\left(\frac{\lambda_1}{\mu} - 1\right) = \epsilon\sqrt{\frac{\lambda_1}{\mu}},$$
$$0 < -E_m\left(\frac{\lambda_n}{\mu} - 1\right) \leq -E_{m_*}\left(\frac{\lambda_n}{\mu} - 1\right) = \epsilon\sqrt{\frac{\lambda_1}{\mu}}.$$

For $x > -1$ and $\gamma(x) = (\sqrt{1+x}-1)/(\sqrt{1+x}+1)$ (as defined in Theorem 3.1), we have $|\gamma(x)| < 1$. Writing down $\gamma(x)$ as

$$\gamma(x) = 1 - \frac{2}{\sqrt{1+x}+1},$$

it is clear that γ is a monotonically increasing function of x. Similarly, we write $E_m(x)$ as

$$E_m(x) = 2\sqrt{1+x}\left(1 - \frac{1}{1+\gamma^{2m+1}(x)}\right)$$

and conclude that $E_m(x)$ is a monotonically increasing function of x. Furthermore, $E_m(x)$ is negative for $-1 < x < 0$ and positive for $x > 0$, and $E_m(0) = 0$. Therefore,

$$|E_m(x)| \le \max\left\{E_m\left(\frac{\lambda_1}{\mu} - 1\right), -E_m\left(\frac{\lambda_n}{\mu} - 1\right)\right\} \le \epsilon\sqrt{\frac{\lambda_1}{\mu}}$$

for $\lambda_n/\mu - 1 \le x \le \lambda_1/\mu - 1$. This concludes our proof for Theorem 3.2. □

We now summarize our approximation result in the following theorem.

THEOREM 3.3. *Let A be an $n \times n$ symmetric positive definite matrix whose largest and smallest eigenvalues are λ_1 and λ_n, respectively. For μ and m given in* (2.3),

$$||\sqrt{A} - S_m||_2 \le \epsilon ||A||_2,$$

where S_m is the approximation given in (2.4).

Proof. Let $x_k = \lambda_k/\mu - 1$ be the kth eigenvalue of X. We have

$$\left\|\sqrt{I+X} - I - \sum_{j=1}^{m} a_j^{(m)}(I + b_j^{(m)}X)^{-1}X\right\|_2 = \max_{1\le k\le n}\left|\sqrt{I+x_k} - 1 - \sum_{j=1}^{m}\frac{a_j^{(m)}x_k}{1+b_j^{(m)}x_k}\right|$$

$$\le \max_{\lambda_n/\mu \le x+1 \le \lambda_1/\mu}\left|\sqrt{I+x} - 1 - \sum_{j=1}^{m}\frac{a_j^{(m)}x}{1+b_j^{(m)}x}\right| \le \epsilon\sqrt{\frac{\lambda_1}{\mu}}.$$

Therefore,

$$\begin{aligned}||\sqrt{A} - S_m||_2 &= ||Q(\sqrt{\mu}\sqrt{I+X} - \sqrt{\mu}I - R_m)Q^T||_2 \\ &= ||\sqrt{\mu}\sqrt{I+X} - \sqrt{\mu}I - R_m||_2 \le \epsilon\sqrt{\lambda_1} = \epsilon||\sqrt{A}||_2.\end{aligned}$$

This concludes our proof. □

For the given error tolerance ϵ and extreme eigenvalues λ_1, λ_n, we use (2.1) to solve t, then calculate m and μ from (2.3). To provide a good initial guess for t, we develop an asymptotic expansion for small t (i.e., large m). Equation (2.1) can be written as

$$e^{t(\ln\alpha + \ln\beta)} + 1 = \left(1 + \frac{2}{\sqrt{\kappa}-1}\right)(e^{t\ln\alpha} + e^{t\ln\beta}).$$

For small t, we have the following expansion:

$$\begin{aligned}&2 + t(\ln\alpha + \ln\beta) + \frac{t^2}{2}(\ln\alpha + \ln\beta)^2 + \cdots \\ &= \left(1 + \frac{2}{\sqrt{\kappa}-1}\right)\left[2 + t(\ln\alpha + \ln\beta) + \frac{t^2}{2}(\ln^2\alpha + \ln^2\beta) + \cdots\right].\end{aligned}$$

TABLE 3.1
A comparison of exact m_* with the approximation in (3.5).

κ	Exact m_*	Approximate m_*
16	4.727	4.476
81	7.467	7.316
256	9.996	9.890
625	12.407	12.328
1296	14.736	14.673
2401	17.000	16.949
4096	19.209	19.167
6561	21.371	21.335
10^4	23.491	23.460
6.25×10^6	91.029	91.030
10^8	151.890	151.893
1.6×10^9	233.322	233.326

This gives rise to

$$t^2 \ln \alpha \ln \beta \sim \frac{4}{\sqrt{\kappa} - 1}$$

or

$$t \sim \frac{2}{\sqrt{(\kappa^{1/2} - 1) \ln \alpha \ln \beta}}, \tag{3.4}$$

where α and β are given in (2.2). The right-hand side above is a good initial guess for t. It can be used by Newton's method to solve t from (2.1). Since $t = 1/(2m_* + 1)$, we also obtain the following asymptotic formula for m_*:

$$m_* \sim \frac{\sqrt{\kappa^{1/2} - 1}}{4} \sqrt{\ln\left(\frac{2}{\epsilon} - 1\right) \ln\left(1 + \frac{2}{\epsilon\sqrt{\kappa}}\right)} - \frac{1}{2}. \tag{3.5}$$

This formula is very accurate. A less accurate and slightly simpler formula for m is (1.3). In Table 3.1, we compare the exact and approximate values of m_* for a few different values of κ.

Since what we are looking for is the integer m greater than or equal to m_*, it is clear that the approximation (3.5) serves that purpose extremely well. The dependence of m_* on $\kappa = \lambda_1/\lambda_n$ is shown in Figure 3.1. Both the exact and approximate values of m_* are shown, but there is no noticeable difference. The error of the approximation formula (3.5) is shown in Figure 3.2.

Based on the above asymptotic formula for t and the expansion $\alpha^t = e^{t \ln \alpha} \approx 1 + t \ln \alpha + \cdots$, we have

$$\mu = \lambda_1 \left(\frac{\alpha^t - 1}{\alpha^t + 1}\right)^2 \sim \lambda_1 \left(\frac{t \ln \alpha}{2}\right)^2.$$

Therefore, for large κ, the scaling parameter μ is given by

$$\mu \sim \frac{\ln \alpha}{\ln \beta} \sqrt{\lambda_1 \lambda_n}. \tag{3.6}$$

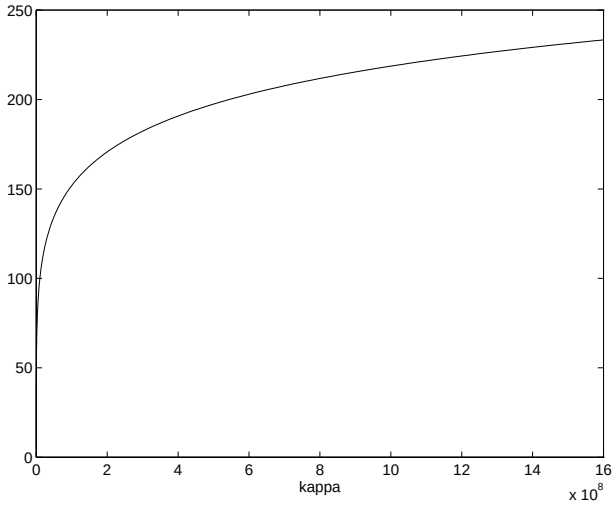


FIG. 3.1. *The dependence of m_* on κ.*

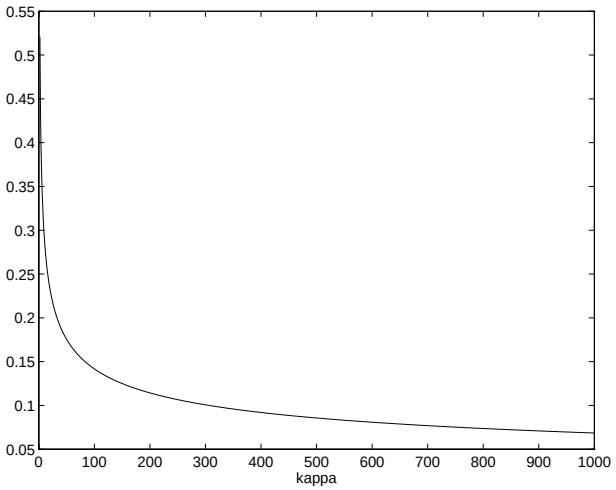


FIG. 3.2. *The difference between exact and approximate values of m_*.*

4. Numerical examples. The method outlined in section 2 has been implemented in a FORTRAN program. It calculates an approximation S_m to $\sqrt{A}$ for a desired relative accuracy ϵ, where A is the given real symmetric positive definite matrix. To demonstrate the accuracy and efficiency of our method, we compare our results with the "exact" solution obtained from a direct calculation of the spectral decomposition (1.2). Relative errors in the different matrix norms are calculated. We denote these errors in Frobenius norm, 1-norm, and 2-norm by e_f, e_1, and e_2, respectively. Namely,

$$e_f = \frac{||\sqrt{A} - S_m||_f}{||\sqrt{A}||_f}, \quad e_1 = \frac{||\sqrt{A} - S_m||_1}{||\sqrt{A}||_1}, \quad e_2 = \frac{||\sqrt{A} - S_m||_2}{||\sqrt{A}||_2}.$$

The following calculations are performed on a SUN Ultra 1 (model 170) workstation.

TABLE 4.1
Relative errors for approximating the square root of Example 1.

n	m	$\kappa(A)$	e_f	e_1	e_2
10	6	29.4410	4.43E-06	7.13E-06	5.41E-06
100	7	41.1412	1.84E-06	6.09E-06	2.93E-06
200	7	41.8252	2.05E-06	9.38E-06	3.19E-06
300	7	42.0545	2.13E-06	1.31E-05	3.50E-06
400	7	42.1694	2.34E-06	1.80E-05	3.94E-06
500	7	42.2385	2.39E-06	2.23E-05	4.94E-06

TABLE 4.2
Relative errors of different methods for $n = 300$.

Method	e_f	e_1	e_2
Padé: $\epsilon = 10^{-6}$, $m = 8$	1.15E-06	1.40E-05	2.80E-06
Padé: $\epsilon = 10^{-7}$, $m = 10$	1.11E-06	1.19E-05	2.82E-06
Spectral decomposition	2.75E-06	2.20E-05	4.56E-06

Example 1. The (i, j) entry of the $n \times n$ matrix is

$$a_{ij} = \frac{1}{2 + (i - j)^2}.$$

We calculate the square root of this matrix for the desired error tolerance $\epsilon = 10^{-5}$ in single precision, then compare the result with the double precision "exact" solution obtained by the spectral decomposition method. This is a well-conditioned matrix and the condition number $\kappa(A)$ grows with n very slowly. The integer m to achieve the desired accuracy is quite small. In Table 4.1, we list m, $\kappa(A)$, and the relative errors for various n. We notice that the error in the 2-norm is indeed bounded by $\epsilon = 10^{-5}$. For the relatively large values of n, the accuracy may not be improved by choosing a smaller ϵ in a single precision calculation, due to round-off errors. However, this is consistent with the single precision result obtained from the standard spectral decomposition method. In Table 4.2, we list the relative errors for $\epsilon = 10^{-6}$ and 10^{-7}, together with those errors for the single precision spectral decomposition method. We observe that the results by our method for all three selections of ϵ are actually more accurate than the result obtained from the standard method.

Example 2. This is the coefficient matrix associated with the standard second order finite difference discretization of the Laplacian on a unit square with Dirichlet boundary conditions. The (i, j) entry of this $n \times n$ matrix is given by

$$a_{ij} = \begin{cases} 4 & \text{if } i = j, \\ -1 & \text{if } |i - j| = p, \\ -1 & \text{if } |i - j| = 1 \text{ and } (i + j) \bmod (2p) \neq 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $n = p^2$. The condition number κ of this matrix is known to be proportional to n^2. Since the dominant term for m is $\kappa^{1/4}$ (see equation (1.3)), we expect that m is roughly proportional to p. This is confirmed in Table 4.3 where the relative errors are also listed.

These calculations are performed for the desired accuracy $\epsilon = 10^{-5}$ in single precision; then the results are compared with the double precision "exact" solution obtained from the standard spectral decomposition method which is based on the

TABLE 4.3
Relative errors for approximating the square root of Example 2.

n	m	e_f	e_1	e_2
16	4	4.74E-06	1.21E-05	9.65E-06
36	5	3.79E-06	1.37E-05	9.95E-06
64	6	2.36E-06	1.07E-05	7.08E-06
100	7	1.49E-06	7.75E-06	4.65E-06
144	8	1.02E-06	6.21E-06	3.07E-06
196	8	2.08E-06	1.27E-05	6.67E-06
256	9	1.39E-06	1.14E-05	4.27E-06
324	9	2.26E-06	1.61E-05	6.99E-06
400	10	1.59E-06	1.65E-05	4.05E-06
484	10	2.25E-06	2.41E-05	6.65E-06

TABLE 4.4
Execution times in seconds by the Padé approximation method (T_{new}) and the spectral decomposition method (T_{old}).

Matrix	n	T_{new}	T_{old}
Example 1	100	0.05	0.12
Example 1	200	0.27	0.90
Example 1	300	0.84	2.90
Example 1	400	2.07	7.15
Example 1	500	4.64	14.81
Example 2	100	0.04	0.12
Example 2	196	0.25	0.81
Example 2	289	0.77	2.37
Example 2	400	2.11	6.55
Example 2	484	4.40	13.21

LAPACK routine `xSYEV` for the spectral decomposition (1.1) and a straightforward evaluation of (1.2). When n is not very small, the result obtained for $\epsilon = 10^{-5}$ is about as accurate as the single precision result by the standard method. For $n = 400$, the spectral decomposition method gives rise to a numerical solution whose relative errors are

$$e_f = 2.45\text{E-}06, \quad e_1 = 2.48\text{E-}05, \quad e_2 = 4.77\text{E-}06.$$

Notice that the last column in Table 4.3 is always less than ϵ, as it is proved in Theorem 3.3.

For both examples, a significant reduction in the total execution time is observed when our method is compared with a single precision computation by the spectral decomposition method. The timing results reported in Table 4.4 are obtained on a SUN Ultra 1 (model 170) workstation based on the compiler f77 (version 4.0) from Sun Microsystems. All programs including LAPACK are compiled with the option "-fast."

Our program uses the LAPACK routine `xSYTRD` for the reduction to tridiagonal form based on Householder reflectors. For banded matrices, the reduction step can be based on Givens rotations and we can use the LAPACK routine `xSBTRD` for this purpose. The matrix in Example 2 is sparse and banded; therefore, the total execution time reported in Table 4.4 can be further reduced. The orthogonal matrix Q for the reduction to tridiagonal form is implicitly given as a product of Given rotations and it is later applied to the matrix R_m as in step 5. Since the reduction step is common to both methods, we expect that the required execution time to decrease roughly the same amount for both methods.

5. Conclusion. In this paper, we have developed a new method for calculating the square root of a symmetric positive definite matrix. While the traditional method based on the spectral decomposition requires about $10n^3$ arithmetic operations, our method requires $\frac{10}{3}n^3 + \frac{5}{2}mn^2$ operations. The number m depends on the spectral condition number κ and the desired accuracy, but it is often quite small compared with n, unless the matrix is very ill conditioned. As is demonstrated in the numerical experiments, for many symmetric positive definite matrices, our method is more efficient than the traditional method based on the spectral decomposition.

The program for this new method can be obtained by e-mail from the author at mayylu@cityu.edu.hk.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, SIAM, Philadelphia, PA, 1992.

[2] G. A. Baker, Jr. and P. Graves-Morris, *Padé Approximations*, 2nd ed., Cambridge University Press, Cambridge, UK, 1996.

[3] A. Bamberger, B. Engquist, L. Halpern, and P. Joly, *Higher order paraxial wave equation approximations in heterogeneous media*, SIAM J. Appl. Math., 48 (1988), pp. 129–154.

[4] A. Björck and S. Hammarling, *A Schur method for the square root of a matrix*, Linear Algebra Appl., 52/53 (1983), pp. 127–140.

[5] L. P. Franca, *An algorithm to compute the square root of a* 3×3 *positive definite matrix*, Comput. Math. Appl., 18 (1989), pp. 459–466.

[6] N. J. Higham, *Newton's method for the matrix square root*, Math. Comp., 46 (1986), pp. 537–550.

[7] N. J. Higham, *Computing real square roots of a real matrix*, Linear Algebra Appl., 88/89 (1987), pp. 405–430.

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

[9] W. B. Jones and W. J Thron, *Continued Fractions, Analytic Theory and Applications*, Addison–Wesley, Reading, MA, 1980.

[10] V. B. Larin, *Calculation of the square root of a positive-definite matrix*, J. Comput. Systems Sci. Internat., 30 (1992), pp. 141–145.

[11] V. B. Larin, *Determination of a square root of a positive definite matrix*, Dokl. Akad. Nauk SSSR, 320 (1991), pp. 536–538 (in Russian).

[12] P. Pulay, *An iterative method for the determination of the square root of a positive definite matrix*, Z. Angew. Math. Mech., 46 (1966), p. 151.

[13] B. A. Schmitt, *Krylov approximations for matrix square roots in stiff boundary value problems*, Math. Comp., 58 (1992), pp. 191–212.

[14] D. Walker and C. Hallum, *Pseudoinverses in generalizing Newton's method for obtaining the square root of a symmetric positive semidefinite matrix*, Indust. Math., 34 (1984), pp. 137–146.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 847–860, October 1998

ON HYPERBOLIC TRIANGULARIZATION: STABILITY AND PIVOTING*

MICHAEL STEWART† AND G. W. STEWART‡

Abstract. This paper treats the problem of triangularizing a matrix by hyperbolic Householder transformations. The stability of this method, which finds application in block updating and fast algorithms for Toeplitz-like matrices, has been analyzed only in special cases. Here we give a general analysis which shows that two distinct implementations of the individual transformations are relationally stable. The analysis also shows that pivoting is required for the entire triangularization algorithm to be stable.

Key words. hyperbolic transformation, triangularization, relational stability, pivoting

AMS subject classifications. 15A23, 65F, 65G05

PII. S0895479897319581

1. Introduction.

Let A be a positive definite matrix of order p and let $R^{\mathrm{T}}R$ be its Cholesky factorization. Given an $m \times p$ matrix X, the *Cholesky updating problem* is to compute the Cholesky factorization

$$\hat{R}^{\mathrm{T}}\hat{R} = \hat{A} \equiv A + X^{\mathrm{T}}X$$

from that of A. Since $X^{\mathrm{T}}X$ is positive semidefinite, $\hat{A}$ is positive definite and always has a Cholesky factor.

It is well known that the Cholesky updating problem can be solved by orthogonal triangularization. Specifically, there is an orthogonal matrix Q such that

$$Q^{\mathrm{T}}\begin{pmatrix} R \\ X \end{pmatrix} = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix},$$

where $\hat{R}$ is upper triangular. From the orthogonality of Q, it follows that

$$\hat{R}^{\mathrm{T}}\hat{R} = \begin{pmatrix} R \\ X \end{pmatrix}^{\mathrm{T}} QQ^{\mathrm{T}}\begin{pmatrix} R \\ X \end{pmatrix} = A + X^{\mathrm{T}}X,$$

so that $\hat{R}^{\mathrm{T}}\hat{R}$ is the required Cholesky factorization. The matrix Q is usually generated as a product of Householder transformations or plane rotations. For details see, e.g., [11].

Now let Y be an $n \times p$ matrix. The *Cholesky downdating problem* is to calculate the Cholesky factor $\hat{R}$ of $\hat{A} = A - Y^{\mathrm{T}}Y$ from that of A. The downdating problem is known to be difficult. An obvious problem is that $\hat{A}$ can be indefinite, in which case the problem has no (real) solution. A more subtle problem is that information present

*Received by the editors April 4, 1997; accepted for publication (in revised form) November 26, 1997; published electronically May 7, 1998.

http://www.siam.org/journals/simax/19-4/31958.html

†Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Australian National University, Canberra ACT 0200, Australia. The work of this author was supported in part by the Department of the Air Force under grant F496220-95-1-0525-P00001.

‡Department of Computer Science and Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742 (stewart@cs.umd.edu). The work of this author was supported by National Science Foundation grant CCR-95-03126.

in the original problem may be represented only imperfectly in the Cholesky factor. For more on this see [17, 3].

The solution to the downdating problem may also be cast in terms of an orthogonal transformation. In particular, if

$$Q\begin{pmatrix} R \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{R} \\ Y \end{pmatrix}, \tag{1.1}$$

then $\hat{R}^{\mathrm{T}}\hat{R} = A - Y^{\mathrm{T}}Y$. Thus if Q is chosen so that $\hat{R}$ is triangular, then $\hat{R}$ is the solution to the downdating problem. Computing such an orthogonal transformation is the basis of a class of algorithms — LINPACK-type algorithms — for this problem see [8, 9, 16].

The Cholesky downdating problem can be solved more directly by an analogue of orthogonal triangularization which we will call *hyperbolic triangularization.* Specifically, a *signature matrix* is a diagonal matrix whose diagonal elements are ± 1. We will say that $\breve{Q}$ is *S-orthogonal* if

$$\breve{Q}^{\mathrm{T}} S \breve{Q} = S,$$

or equivalently if

$$\breve{Q} S \breve{Q}^{\mathrm{T}} = S.$$

Let

$$S = \operatorname{diag}(I_p, -I_n)$$

be the signature matrix corresponding to the partition

$$\begin{pmatrix} R \\ Y \end{pmatrix}.$$

Suppose we can determine an S-orthogonal matrix $\breve{Q}$ such that

$$\breve{Q}^{\mathrm{T}}\begin{pmatrix} R \\ Y \end{pmatrix} = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}. \tag{1.2}$$

Then it follows that $\hat{R}^{\mathrm{T}}\hat{R} = A - Y^{\mathrm{T}}Y$, so that $\hat{R}$ is the downdated Cholesky factor. In practice the matrix $\breve{Q}$ is usually computed as a product of *hyperbolic rotations* [10, 6] or *hyperbolic Householder transformations* [15, 14].

In this paper we will be concerned with the mixed updating problem of calculating the Cholesky factor of $A + X^{\mathrm{T}}X - Y^{\mathrm{T}}Y$ from that of A. Although the problem can be treated as an update followed by a downdate, it is natural to treat it on its own terms. The problem has been considered by Cybenko and Berry [7] in connection with fast algorithms for Toeplitz-like matrices and also by Atkinson [1]. Mixed problems arise naturally in computing the hyperbolic singular value decomposition; a description of the decomposition and signal processing applications are in [13].

The major contribution of this paper is a rounding-error analysis of the use of hyperbolic Householder transformations in the mixed problem. Specifically, we will show that if $\hat{R}$ is the computed value of the new Cholesky factor then there is an orthogonal matrix Q such that

$$Q^{\mathrm{T}}\begin{pmatrix} R \\ X \\ 0 \end{pmatrix} = \begin{pmatrix} \hat{R} \\ 0 \\ Y \end{pmatrix} + E, \tag{1.3}$$

where E is suitably small. Since it is not possible to associate E exclusively with the original data R, X, and Y, this result is not backward stability. Instead we call it *relational stability.* In [18] it is shown that relational stability is preserved under repeated updates and downdates. In consequence, if a Cholesky factor in the sequence is well conditioned it will be computed accurately. By way of previous results, relational stability was established in [17] for the LINPACK algorithm, in [5] for the block downdating problem, and in [4] for the algorithm of [6].

The fact that downdating can be cast in terms of both orthogonal and S-orthogonal matrices suggests that there is a close relation between the two classes. In section 2 we will show that there is indeed a general correspondence between orthogonal matrices with nonsingular principle minors and S-orthogonal matrices. In section 3 we will use this correspondence to derive our hyperbolic triangularization algorithm and two implementations of it. In section 4 we will give rounding-error analyses to show that both implementations are relationally stable. This is in marked contrast to hyperbolic rotations, whose natural implementation is not relationally stable. The relational stability result is relative to certain intermediate quantities which can grow, and in section 5 we show how to use pivoting to control this growth. The paper concludes with some applications and numerical examples.

In this introduction we have placed Cholesky factors to the fore to stress updating and the need for relational stability. However, for simplicity of exposition, we can lump R with X. Thus in the sequel we will consider the hyperbolic triangularization problem of determining an S-orthogonal transformation $\breve{Q}$ such that

$$\breve{Q}^{\mathrm{T}} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}, \tag{1.4}$$

where $\hat{R}$ is upper triangular.

2. Orthogonal and *S*-orthogonal transformations. In this section we will establish a relation between orthogonal and S-orthogonal transformations. The relation can be best described in terms of exchange of variables in linear systems. In what follows S will denote a signature matrix of the form

$$S = \operatorname{diag}(I_m, -I_n).$$

Let Q be an orthogonal matrix and consider the partitioned linear system

$$\begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}, \tag{2.1}$$

where Q_{11} is of order m. If Q_{11} is nonsingular, then the equation $Q_{11}b_1 + Q_{12}b_2 = c_1$ may be solved for b_1 and the result substituted in the equation $Q_{21}b_1 + Q_{22}b_2 = c_2$. The result is a linear system

$$\begin{pmatrix} Q_{11}^{-1} & -Q_{11}^{-1}Q_{12} \\ Q_{21}Q_{11}^{-1} & Q_{22} - Q_{21}Q_{11}^{-1}Q_{12} \end{pmatrix} \begin{pmatrix} c_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ c_2 \end{pmatrix}, \tag{2.2}$$

whose matrix we will denote by $\breve{Q}$. Since the operator that generates $\breve{Q}$ from Q represents an exchange of b_1 and c_1, repeating it will return us to the original system. We will denote this operator by $\operatorname{exc}(Q)$.

We are now ready to state the correspondence between orthogonal and S-orthogonal matrices.[1]

THEOREM 2.1. *Let Q be of order $m+n$ and let Q_{11} be its leading principal submatrix of order n. Let $S = \operatorname{diag}(I_m, -I_n)$. If Q is orthogonal with Q_{11} nonsingular, $\breve{Q} = \operatorname{exc}(Q)$ is S-orthogonal. Conversely, if $\breve{Q}$ is S-orthogonal, then $\breve{Q}_{11}$ is nonsingular and $Q = \operatorname{exc}(\breve{Q})$ is orthogonal.*

Proof. Assume first that Q is orthogonal and Q_{11} is nonsingular. Then (2.1) and the orthogonality of Q imply that

$$\|b_1\|^2 + \|b_2\|^2 = \|c_1\|^2 + \|c_2\|^2.$$

Interchanging b_1 and c_1 gives the system (2.2), in which

$$\|c_1\|^2 - \|b_2\|^2 = \|b_1\|^2 - \|c_2\|^2.$$

In other words, with $d = (c_1^{\mathrm{T}} \;\; b_2^{\mathrm{T}})^{\mathrm{T}}$ we have

$$d^{\mathrm{T}} S d = d^{\mathrm{T}} \breve{Q}^{\mathrm{T}} S \breve{Q} d. \tag{2.3}$$

Now it is easy to see that for any vector c_1 and b_2, there are unique vectors b_1 and c_2 satisfying (2.1). Hence (2.3) holds identically in d. By the uniqueness of quadratic forms, this implies that $\breve{Q}$ is S orthogonal.

The converse proceeds similarly, provided we can establish the nonsingularity of $\breve{Q}_{11}$. But by the S-orthogonality of $\breve{Q}$, we have

$$\breve{Q}_{11}^{\mathrm{T}} \breve{Q}_{11} = I + \breve{Q}_{21}^{\mathrm{T}} \breve{Q}_{21},$$

which is the sum of a positive definite matrix and a semidefinite matrix. Thus $\breve{Q}_{11}^{\mathrm{T}} \breve{Q}_{11}$ and hence $\breve{Q}_{11}$ is nonsingular. ∎

It follows that there is a one-one correspondence between orthogonal matrices with nonsingular leading principal submatrices and S-orthogonal matrices. The hypothesis that Q_{11} be nonsingular is not a restriction in downdating applications. For if Q satisfies (1.1), then $Q_{11} R = \hat{R}$. But for the downdating to be well posed A and $A - Y^{\mathrm{T}} Y$ must be positive definite. Hence R, $\hat{R}$, and $Q_{11} = \hat{R} R^{-1}$ must be nonsingular.

The correspondence gives us considerable flexibility in the way we implement downdating procedures. The key operation is to compute the matrix-vector product

$$\begin{pmatrix} \breve{Q}_{11} & \breve{Q}_{12} \\ \breve{Q}_{21} & \breve{Q}_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}. \tag{2.4}$$

Let

$$\begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} \begin{pmatrix} \hat{x} \\ y \end{pmatrix} = \begin{pmatrix} x \\ \hat{y} \end{pmatrix} \tag{2.5}$$

be the corresponding orthogonal system. Then we have the following options.

1. Apply $\breve{Q}$ directly as in (2.4).
2. Compute $\hat{x}$ from (2.4) and then $\hat{y}$ from (2.5).
3. Solve for $\hat{x}$ in (2.5) and then compute $\hat{y}$ from the same system.

[1] Paul Van Dooren has informed us that the result is a folk theorem in circuit theory, although references to the general result seem to be hard to find (see [2] for a special case). The proof is adapted from a communication by Van Dooren.

Similar alternatives exist when the system that results from exchanging y and $\hat{y}$ is available. If both are available there is a fourth alternative.

4. Solve for $\hat{x}$ in (2.5) and solve for $\hat{y}$ in the system that results from exchanging y and $\hat{y}$.

All these alternatives are mathematically equivalent. But depending on their implementations they may differ numerically.

An important special case is when Q_{11} is a scalar. In particular, if $\breve{Q}$ is a hyperbolic rotation — i.e., x and y are scalars — the second and third strategies listed above yield a mixed algorithm, first presented in [6], that is relationally stable. Another case is when only x is a scalar and Q is a Householder transformation. Again the second and third strategies yield a relationally stable algorithm [15, 14].

3. Hyperbolic Householder transformations and hyperbolic triangularization. The correspondence result of the last section provides a natural way to move from elementary orthogonal transformations, which are used in updating, to S-orthogonal equivalents, which are used in downdating. In particular, we will be concerned with the S-orthogonal equivalent of the Householder transformation

$$H = I - uu^{\mathrm{T}},$$

where u is a vector with $\|u\| = \sqrt{2}$. Let

$$u = \begin{pmatrix} u_x \\ u_y \end{pmatrix},$$

where u_x is of dimension m. Since

$$(I - u_x u_x^{\mathrm{T}})^{-1} = I + u_x u_x^{\mathrm{T}}/c, \tag{3.1}$$

where

$$c = 1 - \|u_x\|^2, \tag{3.2}$$

we see that the S-orthogonal transformation corresponding to H is

$$\breve{H} = \begin{pmatrix} I + u_x u_x^{\mathrm{T}}/c & u_x u_y^{\mathrm{T}}/c \\ -u_y u_x^{\mathrm{T}}/c & I - u_y u_y^{\mathrm{T}}/c \end{pmatrix} = I + Suu^{\mathrm{T}}/c.$$

The case $c = 0$ occurs if and only if $I - u_x u_x^{\mathrm{T}}$ is singular. We shall see presently that singularity does not arise in our applications.

Following the natural correspondence between the orthogonal and S-orthogonal cases, we refer to $\breve{H}$ as a *hyperbolic Householder transformation*. The hyperbolic Householder transformations of [15, 14] are symmetric and have the form $S + uu^{\mathrm{T}}/c$. Computationally the two forms are essentially the same. However, the form given here relates more naturally via the correspondence theorem to standard Householder transformations.

The process of computing $\hat{R}$ in (1.4) is analogous to unitary triangularization by Householder transformations. The following is a recursive description. Assume that X is $m \times p$ and Y is $n \times p$. The first step is to partition

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} x_1 & X_2 \\ y_1 & Y_2 \end{pmatrix}$$

and compute a vector u such that

$$(I + Suu^{\mathrm{T}}/c)\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} \rho \mathbf{e}_1 \\ 0 \end{pmatrix}. \tag{3.3}$$

(The scalar ρ is the $(1,1)$ element of the final triangular form $\hat{R}$.)

To compute u, note that by the correspondence theorem, u also satisfies

$$(I - uu^{\mathrm{T}})\begin{pmatrix} \rho \mathbf{e}_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_1 \\ 0 \end{pmatrix}.$$

The orthogonality of $I - uu^{\mathrm{T}}$ implies that $\rho^2 + \|y_1\|^2 = \|x_1\|^2$ or

$$\rho = \pm\sqrt{\|x_1\|^2 - \|y_1\|^2}.$$

For numerical stability we choose the sign of ρ so that

$$\operatorname{sign}(\rho) = -\operatorname{sign}(\xi_1),$$

where ξ_1 is the first component of x_1. It is now easy to verify that

$$u = \nu^{-1}\begin{pmatrix} x_1 - \rho \mathbf{e}_1 \\ -y_1 \end{pmatrix} \equiv \begin{pmatrix} u_x \\ u_y \end{pmatrix}, \tag{3.4}$$

where

$$\nu = \sqrt{\|x\|_1^2 - \rho\xi_1}$$

is taken to make $\|u\| = \sqrt{2}$.

For later reference we make the following observations about u. For the downdating problem to be well posed, we require that

$$\|x_1\| > \|y_1\|.$$

Consequently, because of the choice of sign for ρ, $\|u_1\| > \|u_2\|$. Since $\|u\|^2 = 2$, we have

$$\|u_x\|^2 > 1.$$

This ensures that $-1 \le c < 0$.

Once u has been determined, we compute

$$(I + Suu^{\mathrm{T}}/c)\begin{pmatrix} x_1 & X_2 \\ y_1 & Y_2 \end{pmatrix} = \begin{pmatrix} \hat{X} \\ \hat{Y} \end{pmatrix} \equiv \begin{pmatrix} \rho \mathbf{e}_1 & \hat{X}_2 \\ 0 & \hat{Y}_2 \end{pmatrix}.$$

We can then proceed recursively by triangularizing the submatrix

$$\begin{pmatrix} \hat{X}_2[2{:}m, 2{:}p] \\ \hat{Y}_2[1{:}n, 2{:}p] \end{pmatrix}$$

by a transformation which is S-orthogonal with respect to $\operatorname{diag}(I_{m-1}, I_n)$.

A complete description of our algorithm requires that we specify how we apply our hyperbolic Householder transformations. In section 2 we listed four ways to apply

a hyperbolic transformation. We will now see how these methods appear in the present application.

The first method is the direct application of the transformation

$$\begin{pmatrix} I + u_x u_x^{\mathrm{T}}/c & u_x u_y^{\mathrm{T}}/c \\ -u_y u_x^{\mathrm{T}}/c & I - u_y u_y^{\mathrm{T}}/c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}. \tag{3.5}$$

The second method is to compute $\hat{x}$ from (3.5) and then determine $\hat{y}$ from the orthogonal system

$$\begin{pmatrix} I - u_x u_x^{\mathrm{T}} & -u_x u_x^{\mathrm{T}} \\ -u_y u_x^{\mathrm{T}} & I - u_y u_y^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \hat{x} \\ y \end{pmatrix} = \begin{pmatrix} x \\ \hat{y} \end{pmatrix}. \tag{3.6}$$

These two algorithms are numerically distinct.

The third algorithm to solve for $\hat{x}$ in (3.6) and then determine $\hat{y}$ directly from (3.6). This algorithm is the same as the second, at least when the implementation takes advantage of the structure of the matrices. From (3.1) and (3.6), we have

$$\hat{x} = (I + u_x u_x^{\mathrm{T}}/c)x - (I + u_x u_x^{\mathrm{T}}/c)u_x u_y^{\mathrm{T}} y.$$

After a little simplification this expression becomes

$$\hat{x} = x + \frac{u_x^{\mathrm{T}} x + u_y^{\mathrm{T}} y}{c} u_x,$$

which is just what one obtains from (3.5).

Similarly, when the formulas for fourth method in section 2 are simplified the method is equivalent to (3.5) with its formulas similarly simplified.

Thus the correspondence theorem gives two methods for applying hyperbolic Householder transformations. We will show in the next section that they are both relationally stable. This fact is a little surprising, since the same approach applied to hyperbolic rotations gives two algorithms, only one of which is relationally stable.

The reason for this difference is that a hyperbolic Householder transformation is not generated and applied explicitly. Instead, its application to a vector is represented as a correction to that vector by another vector from the rank-one matrix Suu^{T}/c. If we were to form the matrix and apply it — something economy would keep us from doing in practice — the algorithm would be as unstable as the direct application of a hyperbolic rotation.

4. The error analysis. In this section we are going to establish the relational stability of hyperbolic triangularization by hyperbolic Householder transformations. We will analyze the algorithm based on direct application of the transformations in detail and then indicate how the same methods apply to the mixed application.

We will analyze the first step of the triangularization algorithm. The step is representative of the others. The hyperbolic Householder transformation $\breve{H}$ is generated from the vector $(x_1^{\mathrm{T}}\ y_1^{\mathrm{T}})^{\mathrm{T}}$.

It is important to note that the transformation associated with a step is used in two different ways. First, it is applied to a general column of the current matrix. Second, it is implicitly applied to the column whose trailing elements are set to zero (in our exposition the first column). Consequently, we will prove two results.

First, let

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \mathrm{fl}\left[\breve{H} \begin{pmatrix} x \\ y \end{pmatrix}\right],$$

where fl denotes the computed value. We will show that there is a Householder transformation $\tilde{H}$ such that

$$\tilde{H}\begin{pmatrix}\hat{x}\\ y\end{pmatrix} = \begin{pmatrix}x\\ \hat{y}\end{pmatrix} + f, \tag{4.1}$$

where $\|f\|$ is small compared to $\|x\|$, $\|y\|$, $\|\hat{x}\|$, and $\|\hat{y}\|$. The transformation $\tilde{H}$ does not depend on x, y, $\hat{x}$, or $\hat{y}$.

Second, in the notation of (3.3), we will show that

$$\tilde{H}\begin{pmatrix}x_1\\ 0\end{pmatrix} = \begin{pmatrix}\rho\mathbf{e}_1\\ y_1\end{pmatrix} + g, \tag{4.2}$$

where g is small compared to $\|x_1\|$ and $\|y_1\|$. Here ρ is the computed quantity. The transformation is computed from x_1 and y_1.

We will assume that the reader is familiar with the basics of rounding error analysis (see, e.g., [12]). Computations are assumed to be done in floating-point arithmetic with rounding unit ϵ_{M}.

Since the object of the analysis is to establish relational stability and not to derive detailed error bounds, we will introduce the following notational simplification. For any vector x with components ξ_i, we will denote generically by $\langle x\rangle$ a vector of the form

$$\langle x\rangle = \begin{pmatrix}\xi_1(1+\epsilon_1)\\ \xi_2(1+\epsilon_2)\\ \vdots\\ \xi_m(1+\epsilon_m)\end{pmatrix},$$

where $\epsilon_i = O(\epsilon_{\text{M}})$. In other words $\langle x\rangle$ is x with its components altered by small relative perturbations. The operator $\langle\cdot\rangle$ has a number of obvious properties — for example,

$$\text{fl}(u^{\text{T}}x) = \langle u\rangle^{\text{T}}x = u^{\text{T}}\langle x\rangle.$$

When the distinction is important, we will use subscripts to distinguish different applications of the operator to the same object — e.g., $\langle x\rangle_1$ and $\langle x\rangle_2$.

Turning now to the analysis itself, we will first establish (4.1). The strategy is the following. After defining $\tilde{u}$, which generates $\tilde{H}$, we show that

$$\begin{pmatrix}I+\tilde{u}_x\tilde{u}_x^{\text{T}}/\tilde{c} & \tilde{u}_x\tilde{u}_y^{\text{T}}/\tilde{c}\\ -\tilde{u}_y\tilde{u}_x^{\text{T}}/\tilde{c} & I-\tilde{u}_y\tilde{u}_y^{\text{T}}/\tilde{c}\end{pmatrix}\begin{pmatrix}\langle x\rangle\\ \langle y\rangle\end{pmatrix} = \begin{pmatrix}\hat{x}+\bar{f}_x\\ \hat{y}+\bar{f}_y\end{pmatrix}, \tag{4.3}$$

where $\bar{f}_x$ and $\bar{f}_y$ are suitably small and $\tilde{c} \equiv 1 - \|\tilde{u}_x\|_2^2$. Since $\tilde{u}$ is the generating vector of a Householder transformation, we may exchange variables to get

$$\tilde{H}\begin{pmatrix}\hat{x}+\bar{f}_x\\ \langle y\rangle\end{pmatrix} = \begin{pmatrix}\langle x\rangle\\ \hat{y}+\bar{f}_y\end{pmatrix}.$$

Equivalently,

$$\tilde{H}\begin{pmatrix}\hat{x}\\ y\end{pmatrix} = \begin{pmatrix}x\\ \hat{y}\end{pmatrix} + \begin{pmatrix}\langle x\rangle - x\\ \bar{f}_y\end{pmatrix} - \tilde{H}\begin{pmatrix}\bar{f}_x\\ \langle y\rangle - y\end{pmatrix} \equiv \begin{pmatrix}x\\ \hat{y}\end{pmatrix} + f,$$

which is of the form (4.1). Multiplication by $\tilde{H}$ in this equation does not magnify the error because $\tilde{H}$ is orthogonal.

We begin with the construction of $\tilde{H}$, which will be done by making small relative perturbations in the components of the computed u to give $\tilde{u}$. For the analysis to go through, these perturbations must accomplish two things. First we must have $\|\tilde{u}\| = \sqrt{2}$, and second we must have

$$\tilde{c} \equiv 1 - \|\tilde{u}_x\|^2 = \langle c \rangle. \tag{4.4}$$

Here c is the computed quantity which satisfies

$$c = 1 - \|\langle u_x \rangle\|^2$$

for some appropriate perturbation of u_x.

Now all that prevents u from being normalized are the small relative errors made in dividing by ν in (3.4). Hence we can normalize u by making small relative perturbations in its elements. However, we must take care that the perturbations enforce (4.4). We consider two cases.

First, assume that $\|u_x\|^2 > 1.5$. Then u_x accounts for a substantial proportion of the norm of u, and we can normalize the latter by small relative perturbations in the former. But in this case, c, which is less than -0.5, is insensitive to such perturbations, so that (4.4) is satisfied.

Second, assume that $\|u_x\|^2 \leq 1.5$. In this case, small relative perturbations in u_x may cause large relative deviation of $1 - \|u_x\|^2$ from c. However, there will always be a small relative perturbation $\tilde{u}_x = \langle u_x \rangle$ such that $c = 1 - \|\tilde{u}_x\|^2$. Having determined $\tilde{u}_x$, we may adjust u_y, which is substantial, to normalize $\tilde{u}$.

In either case, we obtain a normalized $\tilde{u} = \langle u \rangle$ that satisfies (4.4).[2]

We turn now to the application of the transformation. An elementary rounding-error analysis which exploits (4.4) and the fact that $\tilde{u} = \langle u \rangle$ gives the following results:

$$\hat{x} = \langle x \rangle_2 + \frac{\tilde{u}_x^{\mathrm{T}} \langle x \rangle_1 + \tilde{u}_y^{\mathrm{T}} \langle y \rangle_1}{\tilde{c}} \langle \tilde{u}_x \rangle \tag{4.5}$$

and

$$\hat{y} = \langle y \rangle_2 + \frac{\tilde{u}_x^{\mathrm{T}} \langle x \rangle_1 + \tilde{u}_y^{\mathrm{T}} \langle y \rangle_1}{\tilde{c}} \langle \tilde{u}_y \rangle.$$

Subtracting $\langle x \rangle_2$ from both sides of (4.5) and taking norms, we get

$$\left\| \frac{\tilde{u}_x^{\mathrm{T}} \langle x \rangle_1 + \tilde{u}_y^{\mathrm{T}} \langle y \rangle_1}{\tilde{c}} \langle \tilde{u}_x \rangle \right\| \leq \|\hat{x}\| + \|\langle x \rangle_2\|.$$

Hence replacing $\langle x \rangle_2$ by $\langle x \rangle_1$ and $\langle \tilde{u}_x \rangle$ by $\tilde{u}_x$, we may write

$$\hat{x} = \langle x \rangle_1 + \frac{\tilde{u}_x^{\mathrm{T}} \langle x \rangle_1 + \tilde{u}_y^{\mathrm{T}} \langle y \rangle_1}{\tilde{c}} \tilde{u}_x - \bar{f}_x, \tag{4.6}$$

where

$$\|\bar{f}_x\| = (\|\hat{x}\| + \|\langle x \rangle_2\|) O(\epsilon_{\mathrm{M}}).$$

[2] An alternative is to ignore the normalization condition and adjust u_x so that (4.4) is satisfied. The resulting $\tilde{H}$ is not orthogonal but is nearly so. The proof proceeds as usual, except that it must be verified that we can pass from (4.3) to (4.1) without increasing the error. At the end we normalize $\tilde{u}$ and absorb the error into f.

Likewise we can show that

$$\hat{y} = \langle y \rangle_1 + \frac{\tilde{u}_x^{\mathrm{T}} \langle x \rangle_1 + \tilde{u}_y^{\mathrm{T}} \langle y \rangle_1}{\tilde{c}} \tilde{u}_y - \bar{f}_y, \tag{4.7}$$

where

$$\|\bar{f}_y\| = (\|\hat{y}\| + \|\langle y \rangle_2\|) O(\epsilon_{\mathrm{M}}).$$

With a slight change in notation, equations (4.6) and (4.7) are precisely (4.3). Moreover $\|f_x\|$ and $\|f_y\|$ are suitably bounded. Hence we have shown that the direct application of a hyperbolic Householder transformation is relationally stable, i.e., (4.1) holds.

We must now show that $\tilde{H}$ introduces zeros into the first column to working accuracy. Another simple rounding-error analysis (which uses the facts that $\|x_1\| \geq |\rho|$ and that ξ_1 and ρ have opposite signs) shows that

$$\tilde{u} = \nu^{-1} \begin{pmatrix} \langle x_1 \rangle - \langle \rho \rangle \mathbf{e}_1 \\ -\langle y_1 \rangle \end{pmatrix},$$

where $\nu = \sqrt{\|x_1\|^2 - \rho\xi_1}$. It then follows that

$$\begin{aligned} \tilde{H} \begin{pmatrix} x_1 \\ 0 \end{pmatrix} &= \begin{pmatrix} x_1 \\ 0 \end{pmatrix} - \frac{1}{\|x_1\|^2 - \rho\xi_1} \begin{pmatrix} (\langle x_1 \rangle - \langle \rho \rangle \mathbf{e}_1)(\langle x_1 \rangle - \langle \rho \rangle \mathbf{e}_1)^{\mathrm{T}} x_1 \\ -\langle y_1 \rangle (\langle x_1 \rangle - \langle \rho \rangle \mathbf{e}_1)^{\mathrm{T}} x_1 \end{pmatrix} \\ &= \begin{pmatrix} x_1 \\ 0 \end{pmatrix} - \frac{1}{\|x_1\|^2 - \rho\xi_1} \begin{pmatrix} (\langle x_1 \rangle - \langle \rho \rangle \mathbf{e}_1)(\langle \|x_1\|^2 \rangle - \langle \rho \rangle \xi_1) \\ -\langle y_1 \rangle (\langle \|x_1\|^2 \rangle - \langle \rho \rangle \xi_1) \end{pmatrix}. \end{aligned}$$

But because $\rho\xi_1 \leq 0$,

$$\frac{\langle \|x_1\|^2 \rangle - \langle \rho \rangle \xi_1}{\|x_1\|^2 - \rho\xi_1} = \langle 1 \rangle,$$

and (4.2) follows directly.

Turning now to the second form of the algorithm, in which we compute $\hat{x}$ as usual and then compute y from the orthogonal system, we note that this is equivalent to a factored computation of the form

$$\begin{pmatrix} I & 0 \\ u_y u_x^{\mathrm{T}} & I - u_y u_y^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} I + u_x u_x^{\mathrm{T}}/c & u_x u_y^{\mathrm{T}}/c \\ 0 & I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix}.$$

The analysis of this algorithm proceeds as above, with the exception that instead of (4.3) we show that

$$\begin{pmatrix} I & 0 \\ I - \tilde{u}_y \tilde{u}_x^{\mathrm{T}} & -\tilde{u}_y \tilde{u}_y^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} I + \tilde{u}_x \tilde{u}_x^{\mathrm{T}}/\tilde{c} & \tilde{u}_x \tilde{u}_y^{\mathrm{T}}/\tilde{c} \\ 0 & I \end{pmatrix} \begin{pmatrix} \langle x \rangle \\ \langle y \rangle \end{pmatrix} = \begin{pmatrix} \hat{x} + \bar{f}_x \\ \hat{y} + \bar{f}_y \end{pmatrix},$$

where $\bar{f}_x$ and $\bar{f}_x$ are suitably small.

5. Pivoting. In the last section we analyzed the first step of the hyperbolic triangularization algorithm. Since relational stability is cast in terms of orthogonal matrices, we may combine the errors from the several steps to give a relational stability result. However, the errors will be small only when compared to the largest of the intermediate quantities formed in the course of the reduction. If these quantities are

large compared to $\|X\|$ and $\|Y\|$, the final result will be less than satisfactory. Unfortunately, numerical experiments show that growth in the intermediate quantities can occur. See section 6.

The source of the difficulty can be seen by repartitioning $\breve{H}$. Specifically, let (with a change in notation)

$$u = \begin{pmatrix} \upsilon \\ u_x \\ u_y \end{pmatrix},$$

so that

$$\breve{H} \begin{pmatrix} \xi \\ x \\ y \end{pmatrix} = \begin{pmatrix} \xi + \frac{\upsilon\xi + u_x^{\mathrm{T}} x + u_y^{\mathrm{T}} y}{c} \upsilon \\ x + \frac{\upsilon\xi + u_x^{\mathrm{T}} x + u_y^{\mathrm{T}} y}{c} u_x \\ y - \frac{\upsilon\xi + u_x^{\mathrm{T}} x + u_y^{\mathrm{T}} y}{c} u_y \end{pmatrix} \equiv \begin{pmatrix} \hat{\xi} \\ \hat{x} \\ \hat{y} \end{pmatrix}. \tag{5.1}$$

Since $|\upsilon|$, $\|u_x\|$, and $\|u_y\|$ are bounded by $\sqrt{2}$, the quantity

$$\sigma = \frac{\upsilon\xi + u_x^{\mathrm{T}} x + u_y^{\mathrm{T}} y}{c}$$

must be large for there to be any significant increase in the size of the transformed vector. But from the first row of (5.1), we find that

$$|\sigma| \leq \frac{|\xi| + |\hat{\xi}|}{|\upsilon|}.$$

Now $|\xi| \leq \|X\|$. Moreover, in the triangularization algorithm, $\hat{\xi}$ is an element of the final matrix $\hat{R}$ and is also bounded by $\|X\|$. Hence for there to be a growth in the transformed vector, the first component υ of u must be small. But

$$\upsilon = \nu^{-1}(\xi_1 + \operatorname{sign}(\xi_1)\sqrt{\|x_1\|^2 - \|y_1\|^2}).$$

Consequently if the first component of x_1 happens to be small, υ can also be small.

The natural cure is to pivot for size in the matrix X. Specifically, a row interchange is made to move the largest element of x_1 into the first position. This insures that

$$|\xi_1| \geq \frac{\|x_1\|}{\sqrt{m}},$$

and υ cannot be inordinately small. In practice this method has proven effective in preventing growth and stabilizing the triangularization process.

It is worth noting that this problem does not occur in the pure downdating problem where $X = R$. In this case, x_1 has only one component and hence $|\upsilon| > 1$. The same is true in each stage of the reduction.

6. Numerical examples. The first example concerns the necessity and effectiveness of pivoting. Specifically, we consider the updating problem

$$\breve{Q}\begin{pmatrix} R \\ x^{\mathrm{T}} \\ y^{\mathrm{T}} \end{pmatrix} \equiv \breve{Q}\begin{pmatrix} \sqrt{\delta} & 1 \\ 0 & 1 \\ 1 & 1 \\ 1-\sqrt{\delta} & 1 \end{pmatrix} = \begin{pmatrix} \hat{R} \\ \hline 0 \\ \hline 0 \end{pmatrix}$$

for various values of the parameter δ. For the unpivoted algorithm, we compute the error

$$E_{\mathrm{UP}} = \hat{R}^{\mathrm{T}}\hat{R} - (R^{\mathrm{T}}R + xx^{\mathrm{T}} - yy^{\mathrm{T}})$$

and the last elements x_2 and y_2 of x and y after a hyperbolic Householder has been computed to zero the first components. For the pivoted reduction we define E_{P} in analogy with E_{UP}. The following are the results.

δ	$\|E_{\mathrm{UP}}\|$	x_2	y_2	$\|E_{\mathrm{P}}\|$
1.0e−03	1.2e−15	−3.4e+00	−3.3e+00	6.9e−16
1.0e−05	6.7e−14	−1.2e+01	−1.2e+01	6.7e−16
1.0e−07	6.6e−13	−3.9e+01	−3.9e+01	3.9e−17
1.0e−09	1.1e−12	−1.3e+02	−1.3e+02	1.0e−15
1.0e−11	2.0e−12	−4.0e+02	−4.0e+02	5.2e−16
1.0e−13	1.1e−09	−1.3e+03	−1.3e+03	5.6e−16
1.0e−15	3.8e−09	−4.0e+03	−4.0e+03	9.3e−16

For $\delta = 1.0\mathrm{e}{-15}$, we have $\kappa_2(\hat{R}) = 5.6\mathrm{e}{+03}$, which is the worst case. Thus for each value of δ, we have a tractable problem. But in the absence of pivoting there is a significant growth in the intermediate quantities x_2 and y_2, which causes a significant loss in relational stability. The pivoted algorithm has no such problem. These results have been confirmed by experiments with larger, unstructured problems.

We next consider an example of the simple downdating problem which illustrates the difference between hyperbolic rotations and 2×2 hyperbolic Householder transformations. The elementary hyperbolic rotation

$$\breve{Q} = \frac{1}{c}\begin{pmatrix} 1 & s \\ s & 1 \end{pmatrix}$$

for $-1 < c \leq 0$ may be written in hyperbolic Householder form

$$\breve{Q} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} + \frac{1}{c}\begin{pmatrix} \sqrt{1-c} \\ \dfrac{s}{\sqrt{1-c}} \end{pmatrix}\begin{pmatrix} \sqrt{1-c} & \dfrac{s}{\sqrt{1-c}} \end{pmatrix}.$$

The direct application of a hyperbolic rotation is known to be unstable, yet the results of this paper show that the application of the same transformation in the form of a hyperbolic Householder transformation will be relationally stable.

Consider the computation of $\breve{Q}$ such that

$$\breve{Q}\begin{pmatrix} R \\ y^{\mathrm{T}} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1-\delta & 1+\sqrt{\delta} \end{pmatrix} = \begin{pmatrix} \hat{R} \\ 0 \end{pmatrix}.$$

For $\delta < 0.1$ the matrix $R^{\mathrm{T}}R - yy^{\mathrm{T}}$ is positive definite, so that this downdating problem is well posed. We computed a downdated $\hat{R}$ using both hyperbolic rotations and hyperbolic Householders transformations to obtain errors E_{R} and E_{H}. The errors for different values of δ were as follows.

δ	$\|E_{\mathrm{R}}\|$	$\|E_{\mathrm{H}}\|$
1.0e−03	1.1e−15	2.0e−16
1.0e−05	2.7e−14	1.2e−16
1.0e−07	9.7e−14	6.6e−17
1.0e−09	5.6e−13	3.0e−16

Even in this small example, the superior stability of the hyperbolic Householder form of the transformation is evident.

The final example illustrates the importance of relational stability. Specifically, we performed the following experiment. An ill-conditioned matrix R_0 was generated by computing the R-factor from the QR decomposition of a matrix of standard normal deviates (mean zero and standard deviation one). The leading 2×2 principal submatrix was then replaced by the R-factor of another matrix of normal deviates with standard deviation of 10^{-7}. This is a particularly difficult matrix for two reasons.

1. The condition number of R_0 is a little less than the reciprocal square root of the precision used in the experiments (about 10^{-16}). This degree of ill-conditioning is about as great as any downdating algorithm can tolerate.
2. The ill-conditioning is located at the northeast corner of the matrix and will be propagated by the algorithm throughout the matrix. If, instead, the trailing principal submatrix were small, the ill-conditioning would remain localized.

Two auxiliary matrices R_- and R_+ were generated as follows. For $m, n \geq 2$, two $m \times p$ matrices X and Z and an $n \times p$ matrix Y of standard normal deviates were created. The matrix R_+ was obtained by updating Y into R_0. The matrix R_- was obtained in two stages. First, X was updated into R_0. Second, Z was updated into the result while the first row of X was simultaneously downdated. The matrices R_- and R_+ will, in general, be well conditioned, and R_+ was obtained by a process involving a mixed update and downdate. The following display indicates the relation between the matrices ($\oplus$ indicates updating and $\ominus$ downdating).

$$R_- = (R_0 \oplus X) \oplus Z \ominus X[1,:]qaqR_+ = R_0 \oplus Y.$$

We then took the algorithm through the valley of death by downdating Z and the last $m-1$ rows of X to get a matrix $\tilde{R}_0$, which in exact arithmetic would be R_0, and then updating with Y to get a matrix $\tilde{R}_+$, which in exact arithmetic would be R_+. The following is the result of ten runs of the experiment with $p = 20$, $n = 5$, and

$m = 3$.

$\kappa(R_0)$	$\frac{\|\tilde{R}_0 - R_0\|}{\|R_0\|}$	$\kappa(R_+)$	$\frac{\|\tilde{R}_+ - R_+\|}{\|R_+\|}$
2.3e+08	1.7e−02	2.0e+02	4.0e−15
1.3e+08	6.1e−03	1.8e+02	1.6e−15
1.4e+09	3.7e−02	1.4e+03	7.8e−14
6.8e+08	1.0e−02	2.6e+02	1.2e−14
4.8e+08	5.0e−02	2.3e+02	5.1e−15
6.9e+08	2.3e−03	2.8e+02	5.5e−15
1.0e+08	1.3e−03	3.3e+02	1.7e−15
7.4e+08	2.2e−02	1.2e+02	2.5e−15
5.5e+08	5.9e−03	3.1e+02	2.4e−14
3.9e+08	3.1e−03	2.2e+02	5.9e−15

It is seen that in passing from R_- to $\tilde{R}_0$, there is an almost complete loss of accuracy, as predicted by the theory of downdating. On the other hand, almost full accuracy is restored in passing to $\tilde{R}_+$. Such is the power of relational stability to create a silk purse out of a sow's ear.

REFERENCES

[1] E. N. Atkinson, *Computing $A^TA - B^TB = L^TDL$ using generalized hyperbolic transformations*, Linear Algebra Appl., 194 (1993), pp. 135–148.

[2] V. Belovitch, *Classical Network Theory*, Holden Day, San Francisco, CA, 1968.

[3] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, 1996.

[4] A. Bojanczyk, R. P. Brent, P. V. Dooren, and F. de Hoog, *A note on downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 8 (1987), pp. 210–221.

[5] A. W. Bojanczyk and A. O. Steinhardt, *Stability analysis of a Householder-based algorithm for downdating the Cholesky factorization*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1255–1265.

[6] J. M. Chambers, *Regression updating*, J. Amer. Statist. Assoc., 66 (1971), pp. 744–748.

[7] G. Cybenko and M. Berry, *Hyperbolic Householder algorithms for factoring structured matrices*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 499–520.

[8] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, *LINPACK User's Guide*, SIAM, Philadelphia, PA, 1979.

[9] L. Eldén and H. Park, *Block downdating of least squares solutions*, SIAM J. Matrix Anal. Appl., 15 (1992), pp. 1018–1034.

[10] G. H. Golub, *Matrix decompositions and statistical computation*, in Statistical Computation, R. C. Milton and J. A. Nelder, eds., Academic Press, New York, 1969, pp. 365–397.

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

[12] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, PA, 1996.

[13] R. Onn, A. O. Steinhardt, and A. Bojanczyk, *The hyperbolic singular value decomposition and applications*, IEEE Trans. Acoust. Speech Signal Process., 39 (1991), pp. 1575–1588.

[14] C. Rader and A. Steinhardt, *Hyperbolic Householder transforms*, SIAM J. Matrix Anal. Appl., 9 (1988), pp. 269–290. Cited in [3].

[15] C. M. Rader and A. O. Steinhardt, *Hyperbolic Householder transformations*, IEEE Trans. Acoust. Speech Signal Process., 34 (1986), pp. 1589–1602.

[16] M. A. Saunders, *Large-scale Linear Programming Using the Cholesky Factorization*, Technical Report CS252, Computer Science Department, Stanford University, Stanford, CA, 1972. Cited in [3].

[17] G. W. Stewart, *The effects of rounding error on an algorithm for downdating a Cholesky factorization*, J. Inst. Math. Appl., 23 (1979), pp. 203–213.

[18] G. W. Stewart, *On the stability of sequential updates and downdates*, IEEE Trans. Signal Process., 43 (1995), pp. 1643–1648.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 861–871, October 1998

002

APPROXIMATE SEMIDEFINITE MATRICES IN A LINEAR VARIETY*

CHARLES R. JOHNSON† AND PABLO TARAZAGA‡

Abstract. We both characterize and give a convergent algorithm for finding a matrix in a linear variety of matrices that is nearest (in the Frobenius norm) to the positive semidefinite (PSD) matrices. Our motivation is from matrix completions, and in that setting our observations take an especially useful form that we use to bound, and sometimes give closed-form formulae for, the distance from the set of completions to the PSD matrices in terms only of specified data.

Key words. completions, positive semidefinite matrices

AMS subject classifications. 15A57, 15A60, 90C25, 90C30

PII. S0895479896311517

1. Introduction. Given a linear variety L of the $n \times n$ Hermitian matrices H_n, we characterize the matrices in L that are nearest, in the Frobenius norm, to the convex cone Ω_n of positive semidefinite (PSD) matrices. This characterization suggests a simple algorithm whose convergence we prove. In the event that L intersects Ω_n, the algorithm produces a PSD matrix in L. On a very general level (i.e., two convex sets in a vector space) these ideas are not new. However, our motivation is from matrix completion problems, and the specialization of these observations to the problem of a completion nearest to Ω_n takes on a nice and useful form, which we exploit in the final section to bound, and in some cases give closed-form formulas for the distance of the set of completion to Ω_n, in terms only of specified data.

2. Preliminaries and optimality conditions. The subspace of the $n \times n$ matrices M_n consisting of Hermitian matrices is denoted by H_n. As a real vector space, it has dimension n^2, and all considerations herein will take place in this vector space. Results restricted to the $\frac{n(n-1)}{2}$-dimensional subspace of real symmetric matrices are entirely analogous.

Let $\mathcal{S}$ be a (real) subspace of the space of Hermitian matrices H_n, and let $L = A_0 + \mathcal{S}$ for some fixed Hermitian matrix A_0. The orthonormal complement of $\mathcal{S}$ is denoted by $\mathcal{S}^{\perp}$. Let Ω_n be the cone of positive semidefinite matrices. For Hermitian matrices, the usual PSD partial order is defined by

$$X \geq Y \quad \text{if and only if} \quad X - Y \in \Omega_n.$$

The *Frobenius inner product* is defined and denoted as follows:

$$\langle X, Y \rangle_F \stackrel{\text{def}}{=} \operatorname{Tr}(XY^*),$$

then

$$\|X\|_F^2 = \langle X, X \rangle_F.$$

*Received by the editors November 4, 1996; accepted for publication (in revised form) by T. Ando June 30, 1997; published electronically May 14, 1998.
http://www.siam.org/journals/simax/19-4/31151.html

†Department of Mathematics, College of William and Mary, Williamsburg, VA 23187-8795 (crjohnso@math.wm.edu).

‡Department of Mathematics, University of Puerto Rico, Mayaguez, PR 00681-5000 Puerto Rico (tarazaga@jacobi.upr.clu.edu). The research of this author was partially supported by National Science Foundation grants HRD-9450448 and CDA-94117362.

We use the following easily proved, well-known fact:

$$\langle P, Q\rangle_F \geq 0 \quad (P, Q \in \Omega_n). \tag{1}$$

THEOREM 2.1. *For $A \in L$ with Jordan decomposition $A = A_+ + A_-$, the condition*

$$\|A_-\|_F \leq \|T - P\|_F \qquad (T \in L; P \in \Omega_n) \tag{2}$$

is valid if and only if $A_- \in \mathcal{S}^\perp$.

Proof. Suppose first that $A_- \in \mathcal{S}^\perp$. Since any $T \in L$ is in $A + \mathcal{S}$,

$$T - P = A_- + \{A_+ + S - P\} \quad \text{for some } S \in \mathcal{S}.$$

Then apparently,

$$\|T - P\|_F^2 \geq \|A_-\|_F^2 + 2\langle A_-, A_+ + S - P\rangle_F.$$

Since by assumption

$$\langle A_-, A_+\rangle_F = 0 \quad \text{and} \quad \langle A_-, S\rangle_F = 0,$$

and by (1),

$$-\langle A_-, P\rangle_F \geq 0,$$

we can conclude that

$$\|T - P\|_F^2 \geq \|A_-\|_F^2,$$

which leads to (2).

Conversely, suppose that (2) is satisfied. Then for any $S \in \mathcal{S}$, considering $T = A + S, P = A_+$ in (2), we have

$$\|A_- + S\|_F^2 \geq \|A_-\|_F^2.$$

But this is valid for all $S \in \mathcal{S}$ only if $A_- \in \mathcal{S}^\perp$. ∎

The following theorem is very well known.

THEOREM 2.2. *Let $K, -N \in \Omega_n$, and $M = K + N$. Then*

$$\|M_+\|_F \leq \|K\|_F, \quad \|M_-\|_F \leq \|N\|_F.$$

Proof. Since

$$M_+ + M_- = K + N, \quad \text{so that} \quad M_+ - N = K + (-M_-), \tag{3}$$

it follows, from the orthogonality of M_+, M_-, and (1), that by the Schwarz inequality

$$\begin{aligned} \|M_+\|_F^2 = \langle M_+, M_+\rangle_F &\leq \langle M_+, M_+ - N\rangle_F, \\ = \langle M_+, K\rangle_F &\leq \|M_+\|_F \cdot \|K\|_F, \end{aligned}$$

which leads to the first inequality of (3). ∎

We point out in the following example that, although $\|M_+\|_F \leq \|K\|_F$, we do not in general have $M_+ \leq K$.

Example. Let

$$K = \begin{pmatrix} 4 & 2 \\ 2 & 2 \end{pmatrix} \quad \text{and} \quad N = \begin{pmatrix} -1 & -2 \\ -2 & -4 \end{pmatrix};$$

then

$$M = \begin{pmatrix} 3 & 0 \\ 0 & -2 \end{pmatrix} \quad \text{and} \quad M_+ = \begin{pmatrix} 3 & 0 \\ 0 & 0 \end{pmatrix}$$

so that

$$K - M_+ = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix},$$

which is not in Ω_n. Thus $M_+ \not\leq K$.

Our interest is in the problem

$$\text{(L)} \qquad \min_{\substack{A \in L \\ B \in \Omega_n}} \|A - B\|_F = \Delta(L).$$

Since L and Ω_n are closed convex sets, it is a standard fact that $\Delta(L)$ is well defined and is attained at at least one pair $A \in L$, $B \in \Omega_n$.

If $d(A, \Omega_n)$ denotes the minimum distance from a particular $A \in H_n$ to Ω_n, we of course have

$$\Delta(L) = \min_{A \in L} d(A, \Omega_n) = \min_{A \in L} \|A_-\|_F.$$

We are interested in both the structure of the minimizing matrices in L, solutions to (L), and the minimum distance, $\Delta(L)$ from L to Ω_n, primarily in the case of completions, and we study the former in this section. This leads to a natural algorithm, discussed in the next section, and to formulae for the minimum distance, discussed in the final section.

COROLLARY 2.3. *Let A be an $n \times n$ partial Hermitian matrix. The completion $\hat{A}$ is nearest to Ω_n if and only if every entry of $\hat{A}_-$ in an unspecified position of A is* 0.

Proof. From Theorem 2.1, $\hat{A}_-$ is the nearest completion to Ω_n if and only if $A_- \in S^\perp$. But S is generated by element of the canonical basis corresponding to the unspecified entries. Then the orthogonal complement $S^\perp$ is generated by elements of that basis with zero entries in the unspecified positions, which completes the proof. ∎

3. Algorithms. In this section, we present an algorithm for the general linear variety and its particular form for the completion case.

Given a matrix $B \in L$, $\|B_-\|_F$ is the distance from B to Ω_n. Thus we want to minimize $\|B_-\|_F$ in order to find the solution to our problem.

If $\{A_1, \ldots, A_m\}$ is an orthonormal basis for S, the orthogonal complement of $\mathcal{S}$, $\mathcal{S}^\perp$ can be described by

$$S^\perp = span\{A_{m+1}, \ldots, A_{n^2}\},$$

in which $A_{m+1}, \ldots, A_{n^2}$ completes $A_1, \ldots, A_m$ to an orthonormal basis of H_n.

ALGORITHM 1.

Given $B_0 \in L$
for $k = 0, 1, 2, \ldots$
$B_k = (B_k)_+ + (B_k)_-$
$\hat{B}_k = \sum_{i=m+1}^{n^2} \langle (B_k)_-, A_i \rangle_F A_i$
$B_{k+1} = (B_k)_+ + \hat{B}_k$

First of all, observe that, because $\hat{B}_k$ is the projection of $(B_k)_-$ onto $S^\perp$, we have the orthogonal decomposition

$$(B_k)_- = \hat{B}_k + ((B_k)_- - \hat{B}_k).$$

Thus $(B_k)_- - \hat{B}_k$ is an element of S, which implies that

$$B_{k+1} = (B_k)_+ + \hat{B}_k = (B_k)_+ + (B_k)_- - ((B_k)_- - \hat{B}_k) = B_k - ((B_k)_- - \hat{B}_k)$$

is in L.

THEOREM 3.1. *Algorithm* 1*, above, is a descent algorithm, i.e.,*

$$\|(B_{k+1})_-\|_F < \|(B_k)_-\|_F,$$

unless the pair $B_k \in L$*,* $(B_k)_+$ *is a solution to problem* (L).

Proof. Clearly

$$\|\hat{B}_k\|_F^2 < \|(B_k)_-\|_F^2,$$

unless $\hat{B}_k = (B_k)_-$ (i.e., B_k, $(B_k)_+$ is a solution), because $\hat{B}_k$ is the projection of $(B_k)_-$ on $\mathcal{S}^\perp$. But then

$$\hat{B}_k = (\hat{B}_k)_+ + (\hat{B}_k)_-;$$

then

$$B_{k+1} = (B_k)_+ + (\hat{B}_k)_+ + (\hat{B}_k)_-,$$

with $(B_k)_+ + (\hat{B}_k)_+$ in Ω_n and $(\hat{B}_k)_- \in -\Omega_n$. But

$$B_{k+1} = (B_{k+1})_+ + (B_{k+1})_-,$$

and by Theorem 2.2,

$$\|(B_{k+1})_-\|_F \le \|(\hat{B}_k)_-\|_F \le \|(\hat{B}_k)\|_F < \|(B_k)_-\|_F,$$

which proves the theorem. ∎

Remark. We note that, because we are minimizing $\|(B_k)_-\|_F^2$, we may view this algorithm as minimizing the 2-norm of the negative eigenvalues.

In the completion case, the fact that orthogonality of A_- to the linear variety implies that the entries of A_- in the unspecified positions of A have to be 0 simplifies the algorithm. In this event, Algorithm 1 becomes Algorithm 2.

ALGORITHM 2.

Given $B_0 \in L$
for $k = 0, 1, 2, \ldots$
$B_k = (B_k)_+ + (B_k)_-$
$\hat{B}_k = \begin{cases} ((B_k)_-)_{ij} & \text{if the i, j entry is specified} \\ 0 & \text{otherwise} \end{cases}$
$B_{k+1} = (B_k)_+ + \hat{B}_k$

After decomposing $B_k = (B_k)_+ + (B_k)_-$, we simply set equal to 0 those entries of $(B_k)_-$ in the unspecified positions and add the resulting matrix to $(B_k)_+$.

4. Minimum distance to the PSD matrices. Here we restrict our attention to the linear variety of all completions of a given $n \times n$ partial Hermitian matrix A, the graph of whose specified entries is G. As usual, we assume that all diagonal entries of A are specified. Our focus is now upon the minimum value of the objective function in problem (L) rather than upon minimizing completions $\hat{A}$. Let

$$\Delta(A) = \min_{\hat{A} \text{ a completion of } A} \|\hat{A}_-\|_F,$$

the (Frobenius) distance from the linear variety of completions of A to Ω_n.

We may always compute $\Delta(A)$ via the methods of the prior section, etc., but we would like to have a closed-form formula for $\Delta(A)$ in terms of the specified entries of A. This is difficult to obtain in general. However, we are able to give a simple formula in certain cases and, more generally, we relate $\Delta(A)$ to certain other natural quantities associated with the problem. In the process, we introduce some concepts and prove some lemmas that may be of independent interest.

An obvious necessary condition for A to have a completion in Ω_n (i.e., $\Delta(A) = 0$) is that A be *partial* PSD ($A[\alpha] \in \Omega_{|\alpha|}$ whenever $A[\alpha]$ is fully specified). So, a natural quantity is how little the data need be modified in order to make it partial PSD. For a conventional Hermitian matrix B, $G(B)$ is the graph of the nonzero (off-diagonal) entries of B. If B is a conventional matrix, the sum $A + B$ simply means the partial matrix whose graph is G and whose specified entries are those of A plus the corresponding entries of B.

$$\delta(A) = \min_{\substack{G(B) \subset G \\ A+B \text{ is partial PSD}}} \|B\|_F.$$

We refer to a matrix B that attains $\delta(A)$ as a *minimal perturbation* (for A). Of course, it may be that no minimal perturbation yields a partial matrix with PSD completions.

LEMMA 4.1. *For any partial Hermitian matrix A,*

$$\delta(A) \leq \Delta(A),$$

with equality if and only if there is a minimal perturbation of A that has a PSD completion.

Proof. If $\hat{A}$ is a solution to problem (L), then $G(\hat{A}_-) \subset G$, by Corollary 2.3, and $\Delta(A) = \|\hat{A}_-\|_F$. But, $\hat{A} - \hat{A}_- = \hat{A}_+ \in \Omega_n$ and so, $A - \hat{A}_-$ is partial PSD. Thus, $\delta(A) \leq \| - \hat{A}_-\|_F = \|\hat{A}_-\|_F = \Delta(A)$. If equality holds in the last inequality, then $B = -\hat{A}_-$ is a minimal perturbation for A for which $A+B$ has a PSD completion. On the other hand, if B is such a minimal perturbation for A, then $\Delta(A) \leq \|B\|_F = \delta(A)$, which means that $\delta(A) = \Delta(A)$. ∎

We note that implicit in the above is the fact that $\Delta(A)$ is also the minimum by which A may be perturbed to make it PSD completable, i.e.,

$$\Delta(A) = \min_{\substack{A+B \text{ has a} \\ \text{PSD completion}}} \|B\|_F,$$

a fact that is not of direct interest to us here.

In case G is chordal (no induced simple cycles of four or more vertices), partial positive semidefiniteness implies PSD completability [2]. Further, if G is not chordal, there exists partial PSD matrices A ($\delta(A) = 0$) that have no PSD completion ($\Delta(A) > 0$). We may state the following theorem.

THEOREM 4.2. *For every undirected graph G and every partial Hermitian matrix A with graph G, we have*

$$\delta(A) \leq \Delta(A).$$

Equality holds for every partial A with graph G if and only if G is chordal.

We note that, for general (nonchordal) graphs, we know of no certain way to compute $\delta(A)$. It would be of interest, under appropriate normalization, to relate the worst case $\Delta(A) - \delta(A)$ to the graph of G. Let A_0 be the completion of A in which all unspecified entries are chosen to be 0. We may, for example, define

$$\Delta(G) = \max_{\|A_0\|_F=1} (\Delta(A) - \delta(A)).$$

Alternate normalizations might be convenient. Of course, $\Delta(G) = 0$ for all chordal graphs, while $\Delta(G) > 0$ for nonchordal graphs. Is the value of $\Delta(G)$ related to a measure of nonchordality?

We next turn to ideas that permit the calculation of $\Delta(A)$ under special circumstances. For the principal submatrix $A[\alpha]$ of an $n \times n$ (partial or conventional) matrix A, we denote by A_α the $n \times n$ matrix that agrees with A at entries (i, j), when $i, j \in \alpha$, and when it is 0 at other entries. Note that $(A_\alpha)_-$ agrees with $(A[\alpha])_-$ in the α positions (and is zero elsewhere), so that $\|(A_\alpha)_-\|_F = \|(A[\alpha])_-\|_F$. For a partial Hermitian matrix A with graph G, we further define

$$(A_G)_- = \sum_\alpha (A_\alpha)_-,$$

in which the sum is over the maximal cliques α of G.

If A is partial PSD, A_α is PSD so that $(A_\alpha)_- = 0$; hence $(A_G)_- = 0$. In general,

$$(A - (A_G)_-)_\alpha \geq A_\alpha - (A_\alpha)_- = (A_\alpha)_+ \geq 0,$$

so that $A - (A_G)_-$ is always partially PSD. Finally, if the set of summands in $(A_G)_-$ is Frobenius orthogonal, we say that the partial Hermitian A is *negative clique orthogonal* (NCO). A is *special* NCO if it is NCO and if $A - (A_G)_-$ has PSD completions. If G is chordal, all NCO partial matrices are special NCO. A simple way for A to be NCO is for distinct cliques α, for which $A[\alpha]$ is *not* PSD, to be nonoverlapping. We call such an A *negative isolated*. It is possible to be NCO without being negative isolated; in this event, overlapping blocks must themselves be orthogonal. We now define the following key, and easily computed, quantity of interest:

$$\gamma(A) = \left(\sum_\alpha \|A[\alpha]_-\|_F^2 \right)^{1/2}.$$

A related quantity is

$$\Gamma(A) = \|(A_G)_-\|_F.$$

It is then a simple observation that the following holds.

LEMMA 4.3. *For a partial Hermitian matrix A, we have*

$$\gamma(A) \leq \Gamma(A),$$

with equality occurring if and only if A is NCO.

Proof. We may calculate

$$\begin{aligned}\Gamma(A)^2 = \langle (A_G)_-, (A_G)_- \rangle_F &= \left\langle \sum_\alpha (A_\alpha)_-, \sum_\alpha (A_\alpha)_- \right\rangle_F \\ &= \sum_\alpha \|(A_\alpha)_-\|_F^2 + 2 \sum_{\alpha \neq \beta} \langle (A_\alpha)_-, (A_\beta)_- \rangle_F \\ &= \gamma(A)^2 + 2 \sum_{\alpha \neq \beta} \langle (A_\alpha)_-, (A_\beta)_- \rangle_F \geq \gamma(A)^2.\end{aligned}$$

The inner product $\langle (A_\alpha)_-, (A_\beta)_- \rangle_F \geq 0$ because $(A_\alpha)_-, (A_\beta)_- \in -\Omega_n$. Thus $\Gamma(A) \geq \gamma(A)$, with equality exactly when $\sum_{\alpha \neq \beta} \langle (A_\alpha)_-, (A_\beta)_- \rangle_F = 0$, or each $\langle (A_\alpha)_-, (A_\beta)_- \rangle_F = 0$, i.e., A is NCO. ∎

Since $-(A_G)_-$ is the perturbation that makes A partial PSD, it is clear that $\delta(A) \leq \Gamma(A)$. Interestingly, although it is generally smaller than $\Gamma(A)$, $\gamma(A)$ is also an upper bound for $\delta(A)$.

A characterization and parametrization of a minimal member of the set $\{C \,;\, C \geq A, B\}$ is found in [1]. In particular we have the following fact.

LEMMA 4.4. *Suppose that $A,\ B\ \in \Omega_n$. Then*

$$\min_{C \geq A,B} \|C\|_F^2 \leq \|A\|_F^2 + \|B\|_F^2,$$

with equality occurring if and only if $\langle A, B \rangle_F = 0$.

Proof. Let $C = A + (B - A)_+$. Since

$$C = A + (B - A)_+ = A + (B - A) - (B - A)_- = B + (A - B)_+,$$

inequality $C \geq A, B$ is immediate. Remark further that

$$C = \frac{(A + B) + |A - B|}{2}.$$

We have

$$\begin{aligned}\|C\|_F^2 &= \|A\|_F^2 + \|(B - A)_+\|_F^2 + 2\langle A, (B - A)_+ \rangle_F \\ &= \|A\|_F^2 + \|B - A\|_F^2 - \|(B - A)_-\|_F^2 + 2\langle A, B - A \rangle_F - 2\langle A, (B - A)_- \rangle_F \\ &= \|A\|_F^2 + \|B\|_F^2 - \|A + (B - A)_-\|_F^2 \\ &= \|A\|_F^2 + \|B\|_F^2 - \|B - (B - A)_+\|_F^2,\end{aligned}$$

which proves the inequality in the assertion.

If the equality occurs in the assertion, it follows from above that

$$A = -(B - A)_-, \qquad B = (B - A)_+,$$

so that

$$A + B = |B - A|,$$

which implies $\langle A, B \rangle_F = 0$.

Conversely if $\langle A, B \rangle_F = 0$, then generally $C \geq A, B$ implies $C \geq A + B$; hence

$$\|C\|_F^2 \geq \langle A + B, A + B \rangle_F = \|A\|_F^2 + \|B\|_F^2,$$

and the equality occurs. □

The above computation shows that

$$\min_{C \geq A,B} \|C\|_F^2 \leq \|A\|_F^2 + \|B\|_F^2 - \frac{1}{4}\|(A+B) - |A-B|\|_F^2.$$

We note that, when A and B are not orthogonal, the inequality of the lemma may be improved with further manipulation, for example, to

$$\|C\|_F^4 \leq (\|A\|_F^2 + \|B\|_F^2)^2 - \langle A, B\rangle_F^2,$$

and better.

The lemma allows us to do the following. Let

$$A = \begin{pmatrix} A_{11} & A_{12} & ? \\ A_{12}^* & A_{22} & A_{23} \\ ? & A_{23}^* & A_{33} \end{pmatrix}$$

be partial Hermitian, with the two maximal specified principal submatrices

$$A_1 = \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^* & A_{22} \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} A_{22} & A_{23} \\ A_{23}^* & A_{33} \end{pmatrix}.$$

Suppose that

$$(A_1)_- = B = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^* & B_{22} \end{pmatrix} \quad \text{and} \quad (A_2)_- = C = \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^* & C_{22} \end{pmatrix},$$

and define

$$B' = \begin{pmatrix} B_{11} & B_{12} & 0 \\ B_{12}^* & B_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad C' = \begin{pmatrix} 0 & 0 & 0 \\ 0 & C_{11} & C_{12} \\ 0 & C_{12}^* & C_{22} \end{pmatrix}.$$

Then $A - (B' + C')$ is partial PSD, but $\|B' + C'\|_F$ may be bigger than $\gamma(A) = (\|B'\|_F^2 + \|C'\|_F^2)^{1/2}$ to the extent that $\|B_{22} + C_{11}\|_F^2$ exceeds $\|B_{22}\|_F^2 + \|C_{11}\|_F^2$. However, the lemma allows us to replace $B_{22} + C_{11}$ with a matrix D_{22}, such that

$$\|D_{22}\|_F^2 \leq \|B_{22}\|_F^2 + \|C_{11}\|_F^2,$$

and $A - D$ is partial PSD, with

$$D = \begin{pmatrix} B_{11} & B_{12} & 0 \\ B_{12}^* & D_{22} & C_{12} \\ 0 & C_{12}^* & C_{22} \end{pmatrix};$$

then $\|D\|_F \leq \gamma(A)$. Simply pick $-D_{22}$ in relation to $-B_{22}$ and $-C_{11}$ as guaranteed by the lemma. Then

$$-\begin{pmatrix} B_{11} & B_{12} \\ B_{12}^* & D_{22} \end{pmatrix} \geq -B \quad \text{and} \quad -\begin{pmatrix} D_{22} & C_{12} \\ C_{12}^* & C_{22} \end{pmatrix} \geq -C,$$

as $-D_{22} \geq -B_{22}$, $-C_{11}$. Since $A_1 - B = (A_1)_+$ and $A_2 - C = (A_2)_+$ are PSD, this guarantees that $A - D$ is partial PSD. Since $\delta(A) \leq \|D\|_F$, we have $\delta(A) \leq \gamma(A)$.

Repeating the method in Lemma 4.4 we can prove the following.

LEMMA 4.5. *For $B_i \geq 0$ $(i = 1, \ldots, N)$ there is C such that*

$$C \geq B_i \ (i = 1, 2, \ldots, N) \quad \textit{and} \quad \|C\|_F^2 \leq \sum_{i=1}^{N} \|B_i\|_F^2.$$

If

$$\min_{C \geq B_i, i=1,\ldots,N} \|C\|_F^2 = \sum_{i=1}^{N} \|B_i\|_F^2,$$

then

$$\langle B_i, B_j \rangle_F = 0 \qquad (i \neq j).$$

THEOREM 4.6. *For any partial Hermitian matrix A,*

$$\delta(A) \leq \gamma(A).$$

If equality holds, then A is NCO.

Proof. For any maximal clique α let

$$B_{(\alpha)} \stackrel{\text{def}}{=} -(A_\alpha)_-.$$

Then there is C such that $C \geq B_{(\alpha)}$ for all maximal cliques α and

$$\|C\|_F^2 \leq \sum_{\alpha} \|B_{(\alpha)}\|_F^2.$$

Since

$$(A + C)_\alpha \geq (A + B_{(\alpha)})_\alpha = (A_\alpha)_+ \geq 0,$$

$A + C$ is partial PSD. Let $\tilde{C}$ be the projection of C to $\mathcal{S}^\perp$. Then $G(\tilde{C}) \subset G$, $A + \tilde{C}$ is again partial PSD, and

$$\delta(A) \leq \|\tilde{C}\|_F \leq \|C\|_F,$$

which proves the first inequality.

Finally, by Lemma 4.5 $\delta(A) = \gamma(A)$ is possible only when

$$\langle B_i, B_j \rangle = 0 \quad (i \neq j);$$

that is, A is NCO. ∎

We suspect that the converse to the equality statement of the above theorem holds, i.e., if A is NCO, then $\delta(A) = \gamma(A)$. The equality is clear when A is negative isolated, because $(A_G)_-$ is obviously a minimal perturbation. However, even when A is special NCO, we do not know a proof.

We may now give a formula for $\Delta(A)$, when A is special NCO.

THEOREM 4.7. *If A is a special NCO partial Hermitian matrix, then*

$$\Delta(A) = \gamma(A).$$

Proof. Let $\hat{A}$ be a PSD completion of $A-(A_G)_-$, and list the maximal cliques of G as $\alpha_1, \ldots, \alpha_k$. First, notice that $\langle (A_G)_-, \hat{A}\rangle = 0$, since $\langle A[\alpha_i]_-, A[\alpha_i] - A[\alpha_i]_-\rangle_F = 0$ and A is NCO. Now, consider the completion $\hat{A} + (A_G)_-$ of A. Since $\hat{A}$ is PSD and $(A_G)_- \in -\Omega_n$, and the two are orthogonal, we have $\hat{A} = (\hat{A} + (A_G)_-)_+$ and $(A_G)_- = (\hat{A} + (A_G)_-)_-$. Since $G((A_G)_-) \subset G$, the graph of the specified entries of A, we have by Corollary 2.3, that $\hat{A} + (A_G)_-$ is a completion of A that is nearest to Ω_n. Thus, $\Delta(A) = \|(A_G)_-\|_F = \Gamma(A) = \gamma(A)$. ∎

In the case of chordal graphs, the situation is simplified.

COROLLARY 4.8. *Let G be a chordal graph and A a partial Hermitian matrix with graph G. Then,*

$$\Delta(A) \leq \gamma(A).$$

Equality occurs if and only if A is NCO.

Proof. Since G is chordal, $\Delta(A) = \delta(A) \leq \gamma(A)$, i.e., the inequality from Theorems 4.2 and 4.6. Also, since G is chordal, A is NCO if and only if A is special NCO. Thus, if A is NCO, equality follows from the previous theorem. On the other hand, if equality holds, then $\delta(A) = \Delta(A) = \gamma(A)$, and A is NCO, by Theorem 4.6. ∎

In case G is chordal and A is NCO with graph G, we then have

$$\Delta(A) = \gamma(A) = \Gamma(A) = \delta(A).$$

Already in the general chordal case (no NCO assumption), it appears difficult to give a formula for $\Delta(A)$. It may be strictly less than $\gamma(A)$, and no simple correction terms seem to remedy the situation.

Examples. Given the partial matrix

$$A = \begin{pmatrix} 1 & 1 & 1 & ? & 1 \\ 1 & 3/2 & 5/2 & 1 & ? \\ 1 & 5/2 & 1 & 3/2 & ? \\ ? & 1 & 3/2 & 1 & 2 \\ 1 & ? & ? & 2 & 1 \end{pmatrix},$$

we can use Algorithm 2 to compute $\Delta(A) = 1.6799$. This partial matrix A has four specified cliques $\alpha_1 = \{1,2,3\}$, $\alpha_2 = \{2,3,4\}$, $\alpha_3 = \{4,5\}$, and $\alpha_4 = \{1,5\}$. Then we can compute the corresponding $A[\alpha_i]_-$ for $i = 1,2,3,4$.

$$(A[\alpha_1])_- = \begin{pmatrix} -0.0018 & -0.0316 & 0.0357 \\ -0.0316 & -0.5555 & 0.6271 \\ 0.0357 & 0.6271 & -0.7078 \end{pmatrix},$$

$$(A[\alpha_2])_- = \begin{pmatrix} -0.4751 & 0.6233 & -0.1939 \\ 0.6233 & -0.8178 & 0.2544 \\ -0.1939 & 0.2544 & -0.0791 \end{pmatrix},$$

$$(A[\alpha_3])_- = \begin{pmatrix} -0.5000 & 0.5000 \\ 0.5000 & -0.5000 \end{pmatrix}, \qquad (A[\alpha_4])_- = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$

which allows us to compute $\gamma(A) = 2.1173$.

On the other hand, we can generate from the $A[\alpha]_-$,

$$(A_G)_- = \begin{pmatrix} -0.0018 & -0.0316 & 0.0357 & 0 & 0 \\ -0.0316 & -1.0306 & 1.2504 & -0.1939 & 0 \\ 0.0357 & 1.2504 & -1.5256 & 0.2544 & 0 \\ 0 & -0.1939 & 0.2544 & -0.5791 & 0.5000 \\ 0 & 0 & 0 & 0.5000 & -0.5000 \end{pmatrix},$$

and then compute $\Gamma(A) = 2.7948$. Then for this example we have

$$\Delta(A) < \gamma(A) < \Gamma(A).$$

We want to illustrate now inequalities in the other sense, this classical example in PSD completion problems, shows a reversed inequality. Given the partial matrix

$$B = \begin{pmatrix} 1 & 1 & ? & -1 \\ 1 & 1 & 1 & ? \\ ? & 1 & 1 & 1 \\ -1 & ? & 1 & 1 \end{pmatrix},$$

it is easy to see that for every specified clique, $(B[\alpha])_- = 0$ which implies $(B_G)_- = 0$. Then $\gamma(B) = \Gamma(B) = 0$. Now using Algorithm 2, we can compute $\Delta(B) = 0.5858$; then for this matrix B we have

$$\gamma(B) = \Gamma(B) < \Delta(B).$$

In the most general setting we know that

$$\delta(A) \leq \Delta(A), \quad \gamma(A), \quad \Gamma(A),$$

and

$$\gamma(A) \leq \Gamma(A),$$

and that $\Delta(A)$ can be both larger (when A is partial PSD without PSD completions) and smaller (as in the chordal non-NCO case) than both $\gamma(A)$ and $\Gamma(A)$.

Acknowledgments. We want to thank Professor T. Ando and the referee for bringing to our attention reference [1], and for their suggestion to improve this paper by including shorter proofs of Theorems 2.2 and 4.7, and Lemma 4.4.

REFERENCES

[1] T. Ando, *Parametrization of minimal points of some convex sets of matrices*, Acta Sci. Math. (Szeged), 57 (1993), pp. 3–10.

[2] R. Grone, C. R. Johnson, E. Sá, and H. Wolkowicz, *Positive definite completions of partial Hermitian matrices*, Linear Algebra Appl., 58 (1984), pp. 109–124.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 872–885, October 1998

003

KRONECKER STRATIFICATION OF THE SPACE OF QUADRUPLES OF MATRICES*

M$^{\underline{A}}$ I. GARCÍA-PLANAS†

Abstract. In this paper, we study the partition of the space of quadruples of matrices according to the set of discrete structural invariants, proving that it is a stratification and that the structural stability under this equivalence relation is a generic property in the space of quadruples of matrices.

We give an application to obtain bifurcation diagrams for some few-parameters families of quadruples of matrices.

Key words. quadruples of matrices, smooth family, versal deformation, transversality, stratification, bifurcation

AMS subject classifications. 15A21, 93B10, 93B35

PII. S089547989528925X

Introduction. We consider the space $\mathcal{M}_{n,m,p}$ of quadruples of matrices $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ corresponding to a time-invariant linear multivariable system

$$\left.\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned}\right\},$$

and in this space we consider the strict equivalence transformation given by the control system interpretation of basis changes in the state space, the input space and output space, and operations of state feedback and output injection.

The partition of the space of quadruples of matrices into orbits under strict equivalence is not a locally finite partition (under small perturbations, the eigenvalues take an infinity of different values in such a way that the quadruples are nonequivalent). Then, it is not a stratification (although the orbits are smooth submanifolds).

We define in the space $\mathcal{M}_{n,m,p}$ a new equivalence relation in the following manner (see subsections 1.2 and 1.3). We call the Kronecker–Segre symbol (KS-symbol) σ of a quadruple of matrices $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}_{n,m,p}$ the set that consist of the collection of its discrete invariants. Two quadruples of matrices have the same KS type if and only if they have the same KS-symbol σ. Note that then they can differ only in the continuous invariants. An equivalence class, called the Kronecker–Segre stratum (KS-stratum), $E(\sigma)$ in $\mathcal{M}_{n,m,p}$, consists of the set of all quadruples of matrices having a given KS-symbol σ. Obviously, equivalence classes are invariant under strict equivalence.

We prove that this partition is in fact a stratification of the space of quadruples of matrices.

After this we analyze the conditions of stability under this new equivalence relation (Theorem 5.3) from which we can deduce the "genericity" of this property (Theorem 5.6). Given a family $\Lambda \longrightarrow \mathcal{M}_{n,m,p}$ of linear systems parametrized over a space Λ, the above stratification defined in $\mathcal{M}_{n,m,p}$ induces a partition in Λ. However, some extra conditions are necessary in order to ensure that this induced partition Λ is also a stratification. Then we have precise knowledge of the local structures in Λ.

*Received by the editors July 18, 1995; accepted for publication (in revised form) by B. Kågstrom May 15, 1997; published electronically May 15, 1998.

http://www.siam.org/journals/simax/19-4/28925.html

†Departament de Matemàtica Aplicada I, E.T.S. d'Enginyers Industrials, Universitat Politècnica de Catalunya, Marqués de Sentmenat 63, 08029-Barcelona, Spain (igarcia@ma1.upc.es).

These conditions are guaranteed if the family is transversal to the stratification. The Thom Transversality Theorem ensures that it occurs "generically" when the stratification of $\mathcal{M}_{n,m,p}$ is Whitney regular. Then we can speak about "generic families," and for these families a stratification of Λ is obtained (it is called a bifurcation diagram).

In the case where the family is versal, the singularities of the stratification of the space of parameters Λ can be examined after the following theorem (see [1] and [12]): *A family $\varphi : \Lambda \longrightarrow \mathcal{M}_{n,m,p}$ is versal in p_0 if and only if it is transversal to the orbit of $\mathcal{G}$ crossing through $\varphi(p_0) = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ at $p_0 \in \Lambda$.*

For square matrices, V.I. Arnold in [1] studied the bifurcation diagrams partitioning $M_n(\mathbf{C})$ with regard to the Segre symbol, and C.G. Gibson, in [7], proved that this stratification verifies the Whitney regular conditions.

In the case of pairs of matrices, M.I. García-Planas studied in [6] the partition of the set $\mathcal{M}_{n,m} = M_n(\mathbf{C}) \times M_{n\times m}(\mathbf{C})$ with regard to the BK-symbol (that is, the set formed by the controllability indices and the Segre symbol (A, B)) and proved that for $m = 1$ this stratification verifies the Whitney regular conditions.

K. Tchon in [13] proved that the induced stratification over the open dense set of $\mathcal{M}_{n,m}$ formed by the completely reachable pairs of matrices is Whitney regular. This case is immediate since the strata are the orbits under strict equivalence.

We will study here the stratification induced over the open dense subsets of $\mathcal{M}_{n,m,p}$, $\mathcal{A}_{n,m,p} = \{\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right); \operatorname{rank} D = \min(m,p)\}$. It is Whitney regular when we consider the special cases with $m = p$, $m = p + 1$, and $m + 1 = p$.

Finally we enumerate the singularities of "bifurcation diagrams" of some few-parameter families defined using the miniversal deformation.

0. Preliminaries.

0.1. For every integer p, we will denote by $M_p(\mathbf{C})$ the space of p-square complex matrices and by $Gl(p;\mathbf{C})$ the *linear group* formed by the invertible matrices of $M_p(\mathbf{C})$. We will denote by $M_{p\times q}(\mathbf{C})$ the space of rectangular complex matrices having p rows and q columns.

We will deal mainly with quadruples of matrices $\begin{smallmatrix} & n & m \\ p & A & B \\ & C & D \end{smallmatrix}$ such that $A \in M_n(\mathbf{C})$, $B \in M_{n\times m}(\mathbf{C})$, $C \in M_{p\times n}(\mathbf{C})$, and $D \in M_{p\times m}(\mathbf{C})$. $\mathcal{M}_{n,m,p}$ will denote the space of such quadruples of matrices; that is to say, $\mathcal{M}_{n,m,p} = M_n(\mathbf{C}) \times M_{n\times m}(\mathbf{C}) \times M_{p\times n}(\mathbf{C}) \times M_{p\times m}(\mathbf{C})$.

0.2. Throughout this paper the term *manifold* is used as an abbreviation for complex differentiable manifold. We recall that if M is a manifold, X, Y are submanifolds of M, and $x \in X \cap Y$; one says that X, Y are *transversal* at x if the tangent spaces at x verify the relation

$$T_xM = T_xX + T_xY.$$

Then, $X \cap Y$ is also a submanifold of M, and its dimension is $\dim X + \dim Y - \dim M$.

In particular, if

$$T_xM = T_xX \oplus T_xY,$$

we say that X, Y are *minitransversal* at x.

0.3. According to [1, Theorem 5.3] and [8, section 2], the space $M_p(\mathbf{C})$ can be partitioned into a finite number of *Segre Strata*, each one formed by the matrices having the same *Segre symbol* (or the same *Jordan type*). Thus, it is the uncountable union of similarity classes, differing only in the values of the distinct eigenvalues. If $A \in M_p(\mathbf{C})$, we denote by $\sigma_S(A)$, $\mathcal{O}_S(A)$, and $E_S(A)$ the Segre symbol, Segre orbit, and Segre stratum of A, respectively.

If $J \in M_p(\mathbf{C})$ is a Jordan matrix, we denote by $\Gamma_S(J)$ the miniversal deformation of J described in [1, section 4]. In fact, it is a linear variety of $M_p(\mathbf{C})$, minitransversal to $\mathcal{O}_S(J)$ at J.

After [6], the space $\mathcal{M}_{n,m} = \{(A,B) \mid A \in M_n(\mathbf{C}), B \in M_{n\times m}(\mathbf{C})\}$ can be partitioned into a finite number of Brunovsky–Kronecker strata, each one formed by the pairs of matrices having the same discrete invariants.

0.4. A stratification Σ of a subset X of a manifold M is a partition of X into submanifolds of M, called the *strata*, which satisfies the *local finiteness condition*, that is to say, every point in X has a neighborhood in M which meets only finitely many strata.

Let X, Y be disjoint submanifolds in $\mathbf{C}^m$ such that $p \in X \cap \overline{Y}$. Y is said to be regular over X at the point p if for any sequence of points $\{x_n\}$ in X and $\{y_n\}$ in Y converging to p and satisfying the two conditions

(i) the sequence of tangent spaces $T_{y_n}Y$ (regarded as linear subspaces in $\mathbf{C}^m = T_{y_n}(\mathbf{C})^m$) converges to a subspace T in the corresponding Grassmannian;
(ii) the sequence of lines $\overline{x_n y_n}$ converges to a line ℓ in the Grasmannian of lines trough the origin in $\mathbf{C}^m$;

one has $\ell \subset T$.

This is called the Whitney regular conditions.

A Whitney stratification Σ of M is a stratification such that for any pair of strata $X, Y \in \Sigma$, Y is Whitney regular over X at the point p, $\forall p \in X$. In this situation, we may also say that the stratification Σ is Whitney regular.

We recall that a stratification Σ verifies the frontier condition if for any strata X, Y with $X \cap \overline{Y} \neq \emptyset$, then $X \subseteq \overline{Y}$.

Now suppose that Σ is a Whitney regular stratification with connected strata; then Σ verifies the frontier condition. Furthermore, if $X \cap \overline{Y} \neq \emptyset$, then $\dim X < \dim Y$, that is to say, the frontier of a stratum is a union of strata of strictly lower dimension.

We refer to [8, pp. 9–16] for the definitions and results which are needed in what follows.

1. The Kronecker stratification. We shall partition $\mathcal{M}_{n,m,p}$ into a finite number of subsets, each one formed by all the quadruples of matrices having the same complete system of discrete invariants. Hence, each one of these subsets is an orbit or an uncountable union of strict equivalence classes, differing only in the values of the eigenvalues.

In section 4 we shall prove that this partition is in fact a constructible regular stratification. So, we refer to it as the Kronecker–Segre stratification of $\mathcal{M}_{n,m,p}$. Often, we will abbreviate Kronecker by K and Kronecker-Segre by KS, e.g., K-canonical form, KS-stratification, etc.

1.1. We recall that two quadruples of matrices $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right)$ of $\mathcal{M}_{n,m,p}$ are called *similar* if there exist $P \in Gl(n;\mathbf{C})$, $V \in Gl(m;\mathbf{C})$, $W \in Gl(p;\mathbf{C})$, $J \in M_{n\times p}(\mathbf{C})$, and $K \in M_{m\times n}(\mathbf{C})$ such that

$$\left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right) = \left(\begin{smallmatrix} P & J \\ 0 & W \end{smallmatrix}\right)\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)\left(\begin{smallmatrix} P^{-1} & 0 \\ K & V \end{smallmatrix}\right).$$

This is an equivalence relation.

We remark that two quadruples $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right)$ are equivalent if and only if the pencils $\left(\begin{smallmatrix} A+\lambda I_n & B \\ C & D \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} A'+\lambda I_n & B' \\ C' & D' \end{smallmatrix}\right)$ are strictly equivalent. We refer to [10], [11], and [14] for a complete system of invariants and for a Kronecker canonical form that we will denote by $\left(\begin{smallmatrix} A_c & B_c \\ C_c & D_c \end{smallmatrix}\right)$.

1.2. With the above notation, the *KS-symbol* σ of the quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is formed by the column minimal indices k_i, row minimal indices l_i, exponents of infinite elementary divisors $m_i + 1$, and the Segre symbol of J $n_{i,j}$:

$$\begin{aligned}\sigma &= (k, l, \mathrm{m}, \sigma(1), \sigma(2), \ldots, \sigma(u)) \\ &= ((k_1, \ldots, k_r), (l_1, \ldots l_s), (m_1, \ldots, m_t), \\ &\quad (n_{1,1}, n_{1,2}, \ldots), (n_{2,1}, n_{2,2}, \ldots), \ldots, (n_{u,1}, n_{u,2}, \ldots)).\end{aligned}$$

We will denote by $d = \operatorname{rank} D$ and $\delta = \sum_{i,j} n_{ij}$ (that is to say, the size of J).

1.3. A *KS-stratum*, $E(\sigma)$, in $\mathcal{M}_{n,m,p}$ consists of all quadruples of matrices having a given KS-symbol σ. We denote by $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ the K-stratum of the quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$. Then, there are only finitely many KS-strata partitioning $\mathcal{M}_{n,m,p}$. A stratum is an orbit if it is formed by quadruples with no continuous invariants. Otherwise, it is an uncountable union of orbits, differing only in the values of the eigenvalues.

We denote by $\Sigma = \cup_\sigma E(\sigma)$ this partition, and it will be called the *KS-stratification.*

2. The orbits.

2.1. We consider the Lie group

$$\mathcal{G} = \left\{ \left(\left(\begin{smallmatrix} P & J \\ 0 & W \end{smallmatrix}\right), \left(\begin{smallmatrix} P^{-1} & 0 \\ K & V \end{smallmatrix}\right) \right) \;\middle|\; \begin{smallmatrix} P \in Gl(n;\mathbf{C}), V \in Gl(m;\mathbf{C}), W \in Gl(p;\mathbf{C}) \\ J \in M_{n\times m}(\mathbf{C}), K \in M_{m\times n}(\mathbf{C}) \end{smallmatrix} \right\}$$

and its action on $\mathcal{M}_{n,m,p}$ according to the formula

$$\begin{aligned} \alpha : \mathcal{G} \times \mathcal{M}_{n,m,p} &\longrightarrow \mathcal{M}_{n,m,p}, \\ \left(\left(\left(\begin{smallmatrix} P & J \\ 0 & W \end{smallmatrix}\right), \left(\begin{smallmatrix} P^{-1} & 0 \\ K & V \end{smallmatrix}\right) \right), \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \right) &\longrightarrow \left(\begin{smallmatrix} P & J \\ 0 & W \end{smallmatrix}\right) \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \left(\begin{smallmatrix} P^{-1} & 0 \\ K & V \end{smallmatrix}\right). \end{aligned}$$

Given a quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}$, we take $\alpha_{\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)} : \mathcal{G} \longrightarrow \mathcal{M}_{n,m,p}$ as the mapping defined by $g \longrightarrow \alpha(g, \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right))$, for all $g \in \mathcal{G}$.

The equivalence class of a quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ with regard to relation 1.1 is just its orbit by this action: $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) = \alpha_{\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)} \mathcal{G}$.

PROPOSITION 2.1. (1) *The orbits* $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *are constructible sets.*

(2) *The orbits* $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *are complex submanifolds of* $\mathcal{M}_{n,m,p}$.

Proof. First we recall that a set is called constructible if it is a finite union of locally closed sets (see [9] for basic properties). For the proof, we need a theorem of Chevalley (see, for example, [9, Theorem 4.4]): the image of a constructible set under a regular rational mapping is constructible. The assertion follows from the fact that the orbit through $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is the image of the constructible set $\mathcal{G}$ under the rational mapping $\alpha_{\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)}$. (Remember that a regular rational mapping is a mapping of a subset of $\mathbf{C}^p$ into $\mathbf{C}^q$ whose components are rational functions with denominators nowhere zero in the domain.)

Since any constructible set has at least one nonsingular point, and taking into account that the orbits satisfy a homogeneity property (that is to say, given two points on one orbit, there is a diffeomorphism of $\mathcal{M}_{n,m,p}$ mapping one point to the other and preserving orbits), we conclude that every point on an orbit is nonsingular, that is to say, an orbit is a nonsingular constructible set, hence a manifold. □

Remark 2.2. The property of homogeneity of the orbits can be used to reduce our study to quadruples of matrices in their K-canonical form.

2.3. Orbits locally look like the group action quotient stabilizer: let $\mathcal{E}st\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ be the stabilizer of $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ under the action α

$$\mathcal{E}st\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) = \{g \in \mathcal{G} \mid \alpha\left(g, \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)\right) = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)\},$$

and let $V \subset \mathcal{G}$ be a submanifold of $\mathcal{G}$ minitransversal to $\mathcal{E}st\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ at the identity element $I = \left(\left(\begin{smallmatrix} I_n & 0 \\ 0 & I_p \end{smallmatrix}\right), \left(\begin{smallmatrix} I_n & 0 \\ 0 & I_m \end{smallmatrix}\right)\right) \in \mathcal{G}$.

Then, the map defined as $\varphi : V \longrightarrow \mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ by $\varphi(g) = \alpha_{\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)}(g)$ gives a local parametrization of $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$.

3. Local triviality along the orbits. As in the method used in [7], the central point in the study of the stratification is its reduction to the study of the intersection of the strata with a submanifold Γ transversal to the orbits.

The key point is the selection of the submanifold Γ, in order to have an appropriate description of its intersection with the KS-stratification (see Lemma 4.1). In our case, we select Γ as the miniversal deformation defined in [5, subsection II.5]. An explicit description of this "orthogonal" miniversal deformation, as well as a "minimal" miniversal deformation deduced from the "orthogonal" one, was obtained by the author jointly with M. D. Magret and presented at the 3rd IMA conference, July 1995. Such forms have been independently derived by J. Berg and H. Kwatny in [2] and A. Edelman, E. Elmroth, and B. Kågström in [4]; they obtain the miniversal deformation of a pencil and we can deduce a miniversal deformation of a quadruple intersecting the miniversal deformation of the pencil associated with a quadruple with the variety of pencils in the form $A + \lambda B$ with $B = \left(\begin{smallmatrix} I_n & 0 \\ 0 & 0 \end{smallmatrix}\right)$.

3.1. One can get all structures by moving in tangent (1st component of α) and transversal directions (2nd component of α) by means of the following decomposition lemma that provides the desired local trivialization along the orbits. It can be proved by means of the inverse function theorem (in a similar way as in [6, Chapter I, subsection I.3.3] for pairs of matrices).

LEMMA 3.1. *Let* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *be a quadruple in* $\mathcal{M}_{n,m,p}$, $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *its orbit, and* $\Gamma = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + F$ *a linear variety minitransversal to* $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$*:*

$$T\mathcal{M}_{n,m,p} = F \oplus T_{\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)}\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right).$$

Let $\varphi : V \longrightarrow \mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *be a local parametrization as in subsection* 2.3. *Then the mapping*

$$\beta : \Gamma \times \varphi(V) \longrightarrow \mathcal{M}_{n,m,p},$$
$$\beta\left(\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right), \left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right)\right) = \alpha\left(\varphi^{-1}\left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right), \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right)\right)$$

is a local diffeomorphism at $\left(\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right), \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)\right)$ *which preserves the orbits.*

3.2. From [5, subsection II.5], the equations which X, Y, Z, and T must satisfy if $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right) \in \Gamma$ are

$$\left.\begin{aligned} [A, X^*] + BY^* - Z^*C = 0 \\ X^*B + Z^*D = 0 \\ Y^*B + T^*D = 0 \\ CX^* + DY^* = 0 \\ CZ^* + DT^* = 0 \end{aligned}\right\}.$$

Solving this system, it is easy to derive other miniversal deformations with more zero entries in the matrices. In our case we will use the following "minimal" miniversal deformation

$$\left(\begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & A_3 & \\ & & & J \end{pmatrix} \quad \begin{pmatrix} B_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & B_2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & C_1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & C_2 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_d \end{pmatrix}\right) + \left(\begin{pmatrix} 0 & 0 & 0 & 0 \\ X_1^2 & 0 & 0 & X_4^2 \\ 0 & 0 & 0 & 0 \\ X_1^4 & X_2^4 & 0 & X_4^4 \end{pmatrix} \quad \begin{pmatrix} 0 & Y_2^1 & 0 & 0 \\ 0 & Y_2^2 & 0 & 0 \\ Y_1^3 & Y_2^3 & 0 & 0 \\ 0 & Y_2^4 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} Z_1^1 & 0 & Z_3^1 & 0 \\ Z_1^2 & Z_2^2 & Z_3^2 & Z_4^2 \\ 0 & 0 & Z_3^3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} T_1^1 & T_2^1 & T_3^1 & 0 \\ T_1^2 & T_2^2 & T_3^2 & 0 \\ T_1^3 & T_2^3 & T_3^3 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\right),$$

where

$$\begin{aligned} A_1 &= \operatorname{diag}(N_1^1, \dots, N_r^1), & B_1 &= \operatorname{diag}(B_1^1, \dots, B_r^1), \\ A_2 &= \operatorname{diag}(N_1^2, \dots, N_s^2), & B_2 &= \operatorname{diag}(B_1^2, \dots, B_t^2), \\ A_3 &= \operatorname{diag}(N_1^3, \dots, N_t^3), & C_1 &= \operatorname{diag}(C_1^1, \dots, C_s^1), \\ A_4 &= \operatorname{diag}(J_1, \dots, J_u), & C_2 &= \operatorname{diag}(C_1^2, \dots, C_t^2). \end{aligned}$$

$$\begin{aligned} N_i^1 &= \begin{pmatrix} 0 & 0 \\ I_{k_i-1} & 0 \end{pmatrix} \in M_{k_i}(\mathbf{C}),\ 1 \le i \le r, & B_i^1 &= \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}^t \in M_{k_i \times 1}(\mathbf{C}),\ 1 \le i \le r, \\ N_i^2 &= \begin{pmatrix} 0 & I_{l_i-1} \\ 0 & 0 \end{pmatrix} \in M_{l_i}(\mathbf{C}),\ 1 \le i \le s, & B_i^2 &= \begin{pmatrix} 0 & \dots & 0 & 1 \end{pmatrix}^t \in M_{m_i \times 1}(\mathbf{C}),\ 1 \le i \le t, \\ N_i^3 &= \begin{pmatrix} 0 & I_{m_i-1} \\ 0 & 0 \end{pmatrix} \in M_{m_i}(\mathbf{C}),\ 1 \le i \le t, & C_i^1 &= \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} \in M_{1 \times l_i}(\mathbf{C}),\ 1 \le i \le s, \\ J_i &= \operatorname{diag}(J_i^1, \dots, J_i^{\alpha_i}), & C_i^2 &= \begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} \in M_{1 \times m_i}(\mathbf{C}),\ 1 \le i \le t, \\ J_i^\nu &= \lambda_i I_{n_{i,\nu}} + \begin{pmatrix} 0 & 0 \\ I_{n_{i,\nu}-1} & 0 \end{pmatrix} \in M_{n_{i,\nu}}(\mathbf{C}), & & 1 \le i \le u, 1 \le \nu \le \alpha_i. \end{aligned}$$

For the commodity and, if the confusion is not possible, the indices $n_{1,1}, \dots, n_{1,\alpha_1}$, $n_{2,1}, \dots, n_{2,\alpha_2}, \dots, n_{u,1}, \dots, n_{u,\alpha_u}$ can be called $n_1, \dots, n_\delta$.

Also,

(i) Y_1^2, Y_4^2, Z_2^2, Z_2^4, T_1^1, T_2^1, T_3^1, T_1^2, T_2^2, T_3^2, T_1^3, T_2^3, and T_3^3 are free.

(ii) If $X_1^2 = \begin{pmatrix} X_{11}^{21} & \dots & X_{1r}^{21} \\ \vdots & & \vdots \\ X_{s1}^{21} & \dots & X_{sr}^{21} \end{pmatrix}$, then $X_{ij}^{21} = \begin{pmatrix} \star & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \star & 0 & \dots & 0 \end{pmatrix}$ if $k_j \le l_i$, and $X_{ij}^{21} = \begin{pmatrix} \star & \dots & \star \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}$ if $k_j > l_i$.

(iii) All elements in X_4^2 are zero, except those in rows $\{l_1, l_1 + l_2, \dots, l_1 + l_2 + \dots + l_s\}$, that are free.

(iv) All elements in X_1^4 are zero, except those in columns $\{k_1, k_1 + k_2, \dots, k_1 + k_2 + \dots + k_r\}$, that are free.

(v) X_4^4 is Arnold's solution.

(vi) If $Y_1^1 = \begin{pmatrix} Y_{11}^{11} & \dots & Y_{1r}^{11} \\ \vdots & & \vdots \\ Y_{r1}^{11} & \dots & Y_{rr}^{11} \end{pmatrix}$, then $Y_{ij}^{11} = 0$ if $k_i \le k_j + 1$, and $Y_{ij}^{11} = \begin{pmatrix} 0 \\ \star \\ \vdots \\ \star \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ if $k_i > k_j + 1$, with $k_i - k_j - 1$ nonzero entries.

(vii) All elements in Y_2^1 are free, except those in rows $\{1, 1 + k_1, \dots, 1 + k_1 + \dots + k_r\}$, that are zero.

(viii) All elements in Y_1^3 are free, except those in rows $\{m_1, m_1+m_2, \dots, m_1+m_2+\cdots+m_t\}$.

(ix) All elements in Y_2^3 are free, except those in rows $\{m_1, m_1+m_2, \dots, m_1+m_2+\cdots+m_t\}$.

(x) All elements in Z_1^1 are zero, except those in columns $\{1, 1+l_1, \dots, 1+l_1+\cdots+l_{s-1}\}$.

(xi) If $Z_2^1 = \begin{pmatrix} z_{11}^{12} & \dots & z_{1r}^{12} \\ \vdots & & \vdots \\ z_{s1}^{12} & \dots & z_{sr}^{12} \end{pmatrix}$, then $Z_{ij}^{12} = 0$ if $l_i + 1 \geq l_j$ and $Z_{ij}^{12} = (\, 0 \star \dots \star 0 \dots 0 \,)$ if $l_i + 1 < l_j$, with $l_j - l_i - 1$ nonzero entries.

(xii) All elements in Z_3^1 are free, except those in columns $\{1, 1+m_1, \dots, 1+m_1+\cdots+m_{t-1}\}$.

(xiii) All elements in Z_2^2 are free, except those in columns $\{1, 1+l_1, \dots, 1+l_1+\cdots+l_{s-1}\}$.

(xiv) All elements in Z_3^2 are free, except those in columns $\{1, 1+m_1, \dots, 1+m_1+\cdots+m_t\}$.

(xv) If $Z_3^3 = \begin{pmatrix} z_{11}^{33} & \dots & z_{1t}^{33} \\ \vdots & & \vdots \\ z_{t1}^{33} & \dots & z_{tt}^{33} \end{pmatrix}$, then $Z_{ij}^{33} = (\, 0 \dots 0 \star \dots \star \,)$ with $m_j - 1$ nonzero entries.

For this particular variety Γ, it is easy to discuss how the KS-symbol of $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right)$ varies according to the values of the entries of $\left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right)$.

In particular, we have the following.

PROPOSITION 3.2. *With the notation in subsection* 3.2 *and, if* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *is in its canonical reduced form then,*

(a) *If* $\left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right) \neq \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, *then* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right)$ *does not belong to* $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$.

(b) $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right)$ *belongs to* $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *if and only if* $Y = 0$, $Z = 0$, $T = 0$, *and partitioning* X *into blocks corresponding to the blocks in* A, $X = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & X_4^4 \end{pmatrix}$ *and* $J + X_{44}$ *has the same Segre symbol as* J.

4. The strata. Now, we can prove (see Theorem 4.2) that the KS-strata are manifolds, and we give their dimension (see Proposition 4.3). As we have said above, first, in Lemma 4.1 we reduce the problem to the intersection of a KS-stratum with the variety Γ in subsection 3.2. Then, the result follows from the description of this intersection in subsection 3.2, and the fact that the Segre strata are also manifolds.

Other properties of the KS-strata such as constructible and connected are presented in Proposition 4.4 and subsections 4.5 and 4.6.

LEMMA 4.1. *Let* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}_{n,m,p}$, $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *be its orbit,* $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *its stratum, and* Γ *as in Lemma* 3.1. *Then, in a neighborhood of* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *is a submanifold of* $\mathcal{M}_{n,m,p}$ *if and only if* $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \cap \Gamma$ *is a submanifold of* Γ.

Proof. Let us assume that $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is regular at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$. Since Γ is transversal to $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, it is also transversal to $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$. Hence, $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \cap \Gamma$ is regular at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$.

Conversely, let us assume that $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \cap \Gamma$ is regular at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$. According to 3.1, we have

$$E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) = \beta\left(\left(E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \cap \Gamma\right) \times \varphi(V)\right)$$

locally at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$. Therefore, $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is regular at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$. □

4.2. Finally, taking into account the particular form of $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \cap \Gamma$ in 3.2, we have the following theorem.

THEOREM 4.2. *Any KS-stratum is a submanifold of* $\mathcal{M}_{n,m,p}$.

Proof. The proof is obvious in the case when the quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ has no continuous invariants because $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) = \mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$. So, let $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}_{n,m,p}$ be a quadruple with continuous invariants, $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ its orbit, and $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ its KS-stratum. We must prove that $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is regular at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$.

Because of Remark 2.2, we can assume that $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is in its KS-reduced form. By Lemma 4.1 it is sufficient to prove that $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \cap \Gamma$ is regular at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, where Γ is the particular variety in subsection 3.2. Then it follows that $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \cap \Gamma$ is formed by the quadruples of the form

$$\left(\begin{smallmatrix} A' & B \\ C & D \end{smallmatrix}\right), \quad \text{with} \quad A' = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & A_3 & \\ & & & J+X_{44} \end{pmatrix},$$

such that $J + X_{44}$ has the same Segre symbol as J, or equivalently, such that $J + X_{44}$ belongs to the Segre stratum $E_S(J)$ of J.

Therefore, the mapping $\phi : M_\delta(\mathbf{C}) \longrightarrow \mathcal{M}_{n,m,p}$ defined by

$$\phi(M) = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right),$$

with $B = B_c$, $C = C_c$, $D = D_c$, and $A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & A_3 & \\ & & & M \end{pmatrix}$ is an embedding such that

$$\phi(E_S(J) \cap \Gamma_S(J)) = E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \cap \Gamma.$$

(Note that $\phi(J) = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$.)

C. G. Gibson in [7, subsection 2.4] proved that the Segre strata are regular. Hence $E_S(J) \cap \Gamma_S(J)$ is regular at J, (we recall that $\Gamma_S(J)$ is a linear variety transversal to the Segre orbit of J, and hence also transversal to $E_S(J)$ at J), and the proof is completed. □

PROPOSITION 4.3. *Let* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *be a quadruple in* $\mathcal{M}_{n,m,p}$, $\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *be its orbit, and* $E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ *its stratum. Then,*

$$\dim E\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) = u + \dim \mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right),$$

where u *is the number of distinct eigenvalues of* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$.

Proof. It is sufficient to bear in mind that, in the above proof, $\dim(E_S(J) \cap \Gamma_S(J)) = u$ (see [1, subsection 5.5]). □

PROPOSITION 4.4. *The KS-strata are constructible sets.*

Proof. Let $E(\sigma)$ be the KS-stratum corresponding to the KS-symbol

$$\sigma = (\varepsilon, \eta, d, \sigma(1), \ldots, \sigma(u)).$$

Let us consider the set $\mathbf{C}^{(u)} = \{(\lambda_1, \ldots, \lambda_u) \mid \lambda_i \neq \lambda_j \text{ if } i \neq j\} \subset \mathbf{C}^u$. For each $(\lambda_1, \ldots, \lambda_u) \in \mathbf{C}^{(u)}$, let $\mathrm{K}(\sigma; (\lambda_1, \ldots, \lambda_u))$ be the K-matrix of $E(\sigma)$ with eigenvalues $\lambda_1, \ldots, \lambda_u$. Finally, let us consider the mapping $\psi : \mathcal{G} \times \mathbf{C}^{(u)} \longrightarrow \mathcal{M}_{n,m,p}$ defined by

$$\psi(g, (\lambda_1, \ldots, \lambda_u)) = \alpha(g, K(\sigma; (\lambda_1, \ldots, \lambda_u))).$$

Obviously, $\mathcal{G} \times \mathbf{C}^{(u)}$ is a constructible set, ψ is a rational map, and $\psi(\mathcal{G} \times \mathbf{C}^{(u)}) = E(\sigma)$, so that, according to the Chevalley theorem, $E(\sigma)$ is a constructible set. □

4.5. With the notations of the above proof, since ψ is continuous and $\mathcal{G} \times \mathbf{C}^{(u)}$ is connected, we have the following.

PROPOSITION 4.5. *The KS-strata are connected sets.*

4.6. Obviously, the action of $\mathbf{C}$ in $\mathcal{M}_{n,m,p}$ defined by $(\lambda, \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)) \longrightarrow \left(\begin{smallmatrix} A+\lambda I_n & B \\ C & D \end{smallmatrix}\right)$ preserves the KS-strata.

PROPOSITION 4.6. *Let $E(\sigma)$ be a KS-stratum, $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in E(\sigma)$, and $\lambda \in \mathbf{C}$. Then,*

$$\left(\begin{smallmatrix} A+\lambda I_n & B \\ C & D \end{smallmatrix}\right) \in E(\sigma).$$

5. Structural stability. In [5], the author jointly with J. Ferrer studied the structural stability of a quadruple of matrices under the equivalence relation defined in subsection 1.1. J. Demmel and A. Edelman in [3] list in a parallel way the generic structures for nonsquare pencils. They come very close to treating the square case, observing that each eigenvalue considered only as "other than zero or infinity" reduces the codimension of the orbit by one. However, we remark that we cannot deduce the stability of the quadruples by means of the stability of pencils, because a quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ as a pencil $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \lambda \left(\begin{smallmatrix} I_n & 0 \\ 0 & 0 \end{smallmatrix}\right)$ is never structurally stable (it is not equivalent to $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \lambda \left(\begin{smallmatrix} I_n & 0 \\ 0 & D_1 \end{smallmatrix}\right)$, with matrix D_1 having full rank).

Now we are going to identify the generic quadruple, studying the stability under the equivalence relation defined by the strata according to the usual definition.

DEFINITION 5.1. *A quadruple of matrices $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}_{n,m,p}$ is structurally stable if and only if it is an interior point of its stratum.*

PROPOSITION 5.2. *A quadruple is structurally stable if and only if any quadruple in its stratum is structurally stable.*

Proof. The proof is obvious, taking account that $E(\sigma)$ is a connected manifold. □

Taking into account that the structurally stable quadruples of matrices with $m \neq p$ have no finite eigenstructure, its strata coincides with its orbit, then we analyze the stability of strata for quadruples of matrices with $m = p$.

THEOREM 5.3. *A quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}_{n,m,m}$ is structurally stable if and only if* $\operatorname{rank} D = m$ *and the K-canonical form of the quadruple is $\left(\begin{smallmatrix} J & 0 \\ 0 & I \end{smallmatrix}\right)$ and J has distinct eigenvalues.*

Proof. A stratum $E(\sigma)$ is an open set if and only if

$$\dim E(\sigma) = n^2 + nm + np + mp;$$

equivalently, if and only if

$$\dim T\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)^{\perp} = u,$$

where $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is any quadruple in the stratum, and this is verified if and only if the given condition holds. □

Note. The following expression of

$$\dim T\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)^{\perp},$$

in terms of the discrete invariants of $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, was presented at the 3rd IMA conference

in July 1995 by the author jointly with M. D. Magret:

$$\begin{aligned}
&\dim T_{\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)} \mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)^{\perp} \\
&= \sum_{1\le i,j\le r} \max\{0, k_j - k_i - 1\} + \sum_{1\le i\le t} (m_i - 1) \\
&+ \sum_{1\le i\le r} \sum_{1\le j\le s} (k_i + l_j) + r \sum_{1\le i\le \delta} n_i + \sum_{1\le i,j\le s} \max\{0, l_i - l_j - 1\} \\
&+ \sum_{1\le i\le t} (m_i - 1) + \sum_{1\le i,j\le t} (\min\{m_i, m_j\} - 1) + s \sum_{1\le i\le \delta} (n_i) \\
&+ \sum_{1\le i\le u} (n_{i,1} + 3n_{i,2} + 5n_{i,3} + \ldots + (2\alpha_i - 1)n_{i,\alpha_i}) \\
&+ (m - (r + t + d)) \sum_{1\le i\le s} l_i + \left(n - \sum_{1\le i\le r} k_i - \sum_{1\le i\le s} l_i - \sum_{1\le i\le t} m_i \right) (m - (r + t + d)) \\
&+ (p - (s + t + d)) \sum_{1\le i\le r} k_i + \left(n - \sum_{1\le i\le r} k_i - \sum_{1\le i\le s} l_i - \sum_{1\le i\le t} m_i \right) (p - (s + t + d)) \\
&+ (m - (r + t + d)) \sum_{1\le i\le r} (k_i - 1) + (m - (r + t + d)) \sum_{1\le i\le t} (m_i - 1) \\
&+ (p - (s + t + d)) \sum_{1\le i\le t} (l_i - 1) + (p - (s + t + d)) \sum_{1\le i\le t} (m_i - 1) \\
&+ (m - d)(p - d).
\end{aligned}$$

J. Demmel and A. Edelman in [3] give the codimension of the orbit of a pencil $A + \lambda B$ in terms of its discrete invariants. Notice that if we consider the particular pencil where $B = \left(\begin{smallmatrix} I_n & 0 \\ 0 & 0 \end{smallmatrix}\right)$ and counting the codimension of its orbit referred to the variety of pencils $A+\lambda B$ with $B = \left(\begin{smallmatrix} I_n & 0 \\ 0 & 0 \end{smallmatrix}\right)$, the formula presented in [3] coincides with this one.

5.4. Now we are going to deduce conditions for a stratum to be "generic" according to the following definition.

DEFINITION 5.4. *A stratum $E(\sigma)$ is called generic if and only if it is an open dense set and its boundary is the union of strata of lower dimension.*

5.5. Structural stability is a generic property in the space of time-invariant multivariable systems. In fact, when $m \neq p$ generically any two systems of the same dimensions are equivalent.

THEOREM 5.6. *A stratum $E(\sigma)$ is generic if and only if it is structurally stable.*

Proof. Obviously if $E(\sigma)$ is a generic stratum, then it is structurally stable.

Conversely, let $E(\sigma)$ be a structurally stable stratum. To prove that it is a dense set it suffices to consider for any quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}_{n,m,p}$ (that we can take in its K-canonical form) the quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right)$ where entries in the matrices X, Y, Z, T are:

—T is a matrix such that $\operatorname{rank}(D + T) = \min(m, p)$, and $\|T\| < \varepsilon$. We can take such a matrix because the set of rectangular matrices having full rank is an open dense set.

—We fix such a matrix T.

If $m = p$, we take $Y = 0$, $Z = 0$ and the entries in X are such that the matrix

$$A + X - B(D + T)^{-1}C$$

is in the generic stratum of the space of square matrices and $\|X\| < \varepsilon$.

For example, if $A = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$, $C = 0$, and $D = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, then the matrices

$$\left(\begin{smallmatrix} A+X & B \\ C & D+T \end{smallmatrix}\right),$$

where $X = \left(\begin{smallmatrix} \varepsilon_1 & & \\ & \varepsilon_2 & \\ & & \varepsilon_3 \end{smallmatrix}\right)$, $\varepsilon_i - \varepsilon_j \neq 0$ for $i \neq j$, and $T = \left(\begin{smallmatrix} \varepsilon_4 & \\ & 0 \end{smallmatrix}\right)$, $\varepsilon_4 \neq 0$, are in the structurally stable stratum.

If $m > p$, we take $Z = 0$ and let $(I_n, V, W, K, J) \in \mathcal{G}$ be such that

$$\left(\begin{smallmatrix} A+X & B+Y \\ C & D+T \end{smallmatrix}\right) \sim \left(\begin{smallmatrix} A_1 & (B_1\, 0) \\ 0 & (0\, I_p) \end{smallmatrix}\right).$$

Then we take X, Y such that (A_1, B_1) is in the generic stratum of the space of pairs of matrices under block similarity and $\|X\| < \varepsilon$, $\|Y\| < \varepsilon$.

For example, if $A = \left(\begin{smallmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, $C = \left(\begin{smallmatrix} 0 & 0 & 0 \end{smallmatrix}\right)$, and $D = \left(\begin{smallmatrix} 0 & 0 \end{smallmatrix}\right)$, then the matrices

$$\left(\begin{smallmatrix} A+X & B+Y \\ C & D+T \end{smallmatrix}\right),$$

where $X = \left(\begin{smallmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \varepsilon_1 & 0 & 0 \end{smallmatrix}\right)$, $\varepsilon_1 \neq 0$, $Y = 0$, and $T = \left(\begin{smallmatrix} 0 & \varepsilon_2 \end{smallmatrix}\right)$, $\varepsilon_2 \neq 0$, are in the structurally stable stratum.

If $m < p$, we take $Y = 0$ and let $(I_n, V, W, K, J) \in \mathcal{G}$ be such that

$$\left(\begin{smallmatrix} A+X & B \\ C+Z & D+T \end{smallmatrix}\right) \sim \left(\begin{smallmatrix} A_1 & 0 \\ \left(\begin{smallmatrix} C_1 \\ 0 \end{smallmatrix}\right) & \left(\begin{smallmatrix} 0 \\ I_m \end{smallmatrix}\right) \end{smallmatrix}\right).$$

Then we take X, Z such that $\left(\begin{smallmatrix} A_1 \\ C_1 \end{smallmatrix}\right)^t$ is in the generic stratum of the space of pairs of matrices under block similarity, and $\|X\| < \varepsilon$, $\|Z\| < \varepsilon$.

For example, if $A = \left(\begin{smallmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{smallmatrix}\right)$, $B = \left(\begin{smallmatrix} 1 \\ 0 \\ 0 \\ 0 \end{smallmatrix}\right)$, $C = \left(\begin{smallmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{smallmatrix}\right)$, and $D = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$, then the matrices

$$\left(\begin{smallmatrix} A+X & B \\ C+Z & D+T \end{smallmatrix}\right),$$

where $X = \left(\begin{smallmatrix} 0 & 0 & 0 & \varepsilon_1 \\ \varepsilon_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{smallmatrix}\right)$, $\varepsilon_i \neq 0$, $Z = 0$, and $T = \left(\begin{smallmatrix} 0 \\ \varepsilon_3 \end{smallmatrix}\right)$, $\varepsilon_3 \neq 0$, are in the structurally stable stratum.

Clearly $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right) \in E(\sigma)$. Finally, u is a lower boundary for the dimension of $T\mathcal{O}\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)^{\perp}$. □

6. Regularity properties of the KS-stratification. We will study the regularity of the KS-stratification over strata called simple (Proposition 6.2). Since they have a particular homogeneity property (Proposition 6.3), their regularity follows from the Whitney theorem (Proposition 6.4).

In the following we will write as $\mathcal{A}_{n,m,p}$ the space of quadruples of matrices such that the last matrix has full rank, that is to say,

$$\mathcal{A}_{n,m,p} = \left\{\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right);\ \operatorname{rank} D = \min(m, p)\right\} \subset \mathcal{M}_{n,m,p}.$$

Finally we are going to prove that in the spaces $\mathcal{M}_{n,m,m}$, $\mathcal{M}_{n,p+1,p}$, $\mathcal{M}_{n,m,m+1}$ the KS-stratification induced over the open dense sets $\mathcal{A}_{n,m,m} \subset \mathcal{M}_{n,m,m}$, $\mathcal{A}_{n,p+1,p} \subset \mathcal{M}_{n,p+1,p}$, and $\mathcal{A}_{n,m,m+1} \subset \mathcal{M}_{n,m,m+1}$, respectively, satisfies the Whitney regular condition which basically is a fancy continuity condition.

6.1. First, we are going to prove the Whitney regularity condition of the KS-stratification over the so-called *simple* strata.

DEFINITION 6.1. *A quadruple of matrices* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}_{n,m,p}$ *is called simple if it has, at most, one eigenvalue. A stratum* $E(\sigma)$ *is called simple if its elements are simple, that is to say, if* $\sigma = (k, l, \mathrm{m}, \sigma(1))$ *or* $\sigma = (k, l, \mathrm{m})$.

6.2. The simple strata verify a particular homogeneity property (in some sense the converse to the one in subsection 4.6).

PROPOSITION 6.2. *Let* $E(\sigma)$ *be a simple stratum. For any* $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right), \left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right) \in E(\sigma)$, *there exists a diffeomorphism* f *of* $\mathcal{M}_{n,m,p}$, *preserving strata, and such that* $f\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) = \left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right)$.

Proof. If the quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in E(\sigma)$ has no eigenvalues, the proof is trivial. Let λ, λ' be the eigenvalues of $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right)$, respectively. Then because of 4.6 the quadruple $\left(\begin{smallmatrix} A+(\lambda'-\lambda)I_n & B \\ C & D \end{smallmatrix}\right)$ is equivalent to $\left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right)$. Hence, there exists $g_0 \in \mathcal{G}$ such that $\left(\begin{smallmatrix} A' & B' \\ C' & D' \end{smallmatrix}\right) = \alpha\left(g_0, \left(\begin{smallmatrix} A+(\lambda'-\lambda)I_n & B \\ C & D \end{smallmatrix}\right)\right)$. It is straightforward that the mapping

$$f\left(\begin{smallmatrix} X & Y \\ Y & Z \end{smallmatrix}\right) = \alpha\left(g_0, \left(\begin{smallmatrix} X+(\lambda'-\lambda)I_n & Y \\ Z & T \end{smallmatrix}\right)\right)$$

verifies the desired conditions. □

6.3. The Whitney theorem states that any stratum of a constructible locally finite stratification has a Whitney regular point. Hence, any stratum $E(\sigma)$ has a point $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ such that Σ is Whitney regular over $E(\sigma)$ at $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$. In the particular case where $E(\sigma)$ is simple, the above homogeneity property implies that all the points of $E(\sigma)$ are Whitney regular. Therefore, we have the following.

PROPOSITION 6.3. Σ *is Whitney regular over any simple KS-stratum.*

6.4. Finally, we tackle the following.

PROPOSITION 6.4. *The natural stratification in*

(i) $\mathcal{A}_{n,m,m}$,
(ii) $\mathcal{A}_{n,p+1,p}$,
(iii) $\mathcal{A}_{n,m,m+1}$,

induced by the KS-stratification in $\mathcal{M}_{n,m,m}$, $\mathcal{M}_{n,p+1,p}$, $\mathcal{M}_{n,m,m+1}$, *respectively, are Whitney regular.*

Proof. (i) The K-canonical form of square quadruples $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ with D having full rank is $\left(\begin{smallmatrix} J & \\ & I \end{smallmatrix}\right)$. All small perturbations of this quadruple will have K-canonical form $\left(\begin{smallmatrix} J' & \\ & I \end{smallmatrix}\right)$. So, without loss of generality, we can consider only pairs (J, I). C.G. Gibson in [7] shows that V.I. Arnold [1] constructed a Whitney stratification of J. The stratum of I is the whole space of m-square complex matrices, which is trivially Whitney regular. Now, by [8, Theorem 1.2], the product of Whitney stratifications is a Whitney regular stratification of (J, I).

(ii) The canonical form of rectangular quadruples $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ with $m = p + 1$ and D having full rank is $\left(\begin{smallmatrix} A_c & B_c \\ C_c & D_c \end{smallmatrix}\right)$, where $B_c = \left(\begin{smallmatrix} B_{c_1} & 0 \end{smallmatrix}\right)$, $B_{c_1} \in M_{n\times 1}(\mathbf{C})$, $C_c = 0$, $D_c = \left(\begin{smallmatrix} 0 & I_p \end{smallmatrix}\right)$. All small perturbations of this quadruple will have K-canonical form $\left(\begin{smallmatrix} A'_c & B'_c \\ C'_c & D'_c \end{smallmatrix}\right)$, where $B'_c = \left(\begin{smallmatrix} B'_{c_1} & 0 \end{smallmatrix}\right)$, $B'_{c_1} \in M_{n\times 1}(\mathbf{C})$, $C'_c = 0$, $D'_c = \left(\begin{smallmatrix} 0 & I_p \end{smallmatrix}\right)$. So, without loss of generality, we can consider only elements of $\mathcal{M}_{n,1} \times M_p(\mathbf{C})$, where $\mathcal{M}_{n,1}$ is the space of pairs of matrices such that the second matrix is a column matrix. In the space of

pairs of matrices, a stratification was defined in [6], and in the case $\mathcal{M}_{n,1}$ in [6] it is proved to be is Whitney regular. Now it follows as in (i).

(iii) The proof is analogous to (ii). □

7. Bifurcation diagrams. Let $\varphi : \Lambda \longrightarrow \mathcal{M}_{n,m,p}$ be a family of quadruples of matrices transversal to the stratification. Then after [8], the induced partition on Λ is also a stratification, and

$$\operatorname{codim} \varphi^{-1}(E(\sigma)) = \operatorname{codim} E(\sigma).$$

7.1. Since $\mathcal{A}_{n,m,m}$, $\mathcal{A}_{n,p+1,p}$, $\mathcal{A}_{n,m,m+1}$ are endowed with a Whitney stratification, we can make use of the Thom transversality theorem. The set of families of quadruples of matrices transversal to the stratification is open and dense in the spaces $C^\infty(\Lambda, \mathcal{A}_{n,m,m})$, $C^\infty(\Lambda, \mathcal{A}_{n,p+1,1})$, $C^\infty(\Lambda, \mathcal{A}_{n,m,m+1})$, respectively. We call such families "generic."

In these cases the singularities of the bifurcation diagrams of $\varphi^{-1}(\mathcal{A}_{n,m,p})$ ($m = p$, $m = p + 1$, or $p = m + 1$) for a generic family $\varphi : \Lambda \longrightarrow \mathcal{A}_{n,m,p}$ are easily checked.

In the case $m = p$, with the same notation as in subsection 6.4, the bifurcations of (J, I) are exactly those of J. See [1] for discussion and sketches of some simple bifurcations.

In the case $m = p + 1$, also with the same notation as in subsection 6.4, the bifurcations of $\left(\begin{smallmatrix} A_c & B_c \\ C_c & D_c \end{smallmatrix}\right)$ are exactly those of (A_c, B_{c_1}). See [6] for discussion of some bifurcations of few-parameters generic families.

7.2. Taking into account that a miniversal deformation of a quadruple of matrices $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ is a minitransversal family to its orbit, it then follows that it is a transversal family to its stratum. It is possible to describe "bifurcation diagrams" using versal deformations of a quadruple, that is to say, the stratification may be seen directly in the miniversal deformation.

In fact, starting from a miniversal deformation of a quadruple we can obtain a minitransversal family to its stratum (but not transversal to the orbit) in the following manner. Obviously we can reduce the study to the case where the quadruple is in its K-canonical form, and we consider the minimal miniversal deformation Γ defined in subsection 3.2.

If we take X_{44} such that $\operatorname{trace} X_i = 0$, for each diagonal block X_i corresponding to the diagonal blocks in $J_1, \dots J_u$, we obtain a minitransversal family to the stratum $E(\sigma)$.

7.3. Now we are going to use this minitransversal family to the strata in a particular case.

Let $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \in \mathcal{M}_{5,m,p}$ $m \geq p + 2$, where

$$A = \begin{pmatrix} 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&1 \end{pmatrix}, \; B = \begin{pmatrix} 0&0&0&\dots \\ 0&0&0&\dots \\ 1&0&0&\dots \\ 0&1&0&\dots \\ 0&0&0&\dots \end{pmatrix}, \; C = \begin{pmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ \vdots&\vdots&\vdots&\vdots&\vdots \end{pmatrix}, \; D = \begin{pmatrix} 0&0&1&& \\ 0&0&&\ddots& \\ 0&0&&&1 \end{pmatrix}.$$

A minitransversal family to the strata is $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) + \left(\begin{smallmatrix} X & Y \\ Z & T \end{smallmatrix}\right)$, where $X = \begin{pmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ x&0&0&y&0 \end{pmatrix}$, $Y = \begin{pmatrix} 0&0&0&\dots \\ 0&z&0&\dots \\ 0&0&0&\dots \\ 0&0&0&\dots \\ 0&0&0&\dots \end{pmatrix}$, $Z = 0$, $T = 0$.

The singularities of bifurcation diagrams are as follows in a neighborhood of $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, almost all of quadruples are in the generic stratum *formed by* quadruples

in the form $\left(\begin{smallmatrix} A_1 & B_1 \\ C_1 & D_1 \end{smallmatrix}\right)$, where $A_1 = \begin{pmatrix} 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&1 \\ 0&0&0&0&0 \end{pmatrix}$, $B = \begin{pmatrix} 0&0&0&\dots \\ 0&0&0&\dots \\ 1&0&0&\dots \\ 0&0&0&\dots \\ 0&1&0&\dots \end{pmatrix}$, $C = \begin{pmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ \vdots&\vdots&\vdots&\vdots&\vdots \end{pmatrix}$, $D = \begin{pmatrix} 0&0&1 & & \\ 0&0& & \ddots & \\ 0&0& & & 1 \end{pmatrix}$ *situated outside* the hyperbolical paraboloid surface

$$xz + y = 0.$$

Quadruples *situated* in $\{xz + y = 0\} - \{(0, 0, 0)\}$ are in the stratum formed by quadruples in the form $\left(\begin{smallmatrix} A_1 & B_1 \\ C_1 & D_1 \end{smallmatrix}\right)$, where

$$A_1 = \begin{pmatrix} 0&1&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&1&0 \\ 0&0&0&0&0 \\ 0&0&0&0&a \end{pmatrix}, B = \begin{pmatrix} 0&0&0&\dots \\ 1&0&0&\dots \\ 0&0&0&\dots \\ 0&1&0&\dots \\ 0&0&0&\dots \end{pmatrix}, C = \begin{pmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ \vdots&\vdots&\vdots&\vdots&\vdots \end{pmatrix}, D = \begin{pmatrix} 0&0&1 & & \\ 0&0& & \ddots & \\ 0&0& & & 1 \end{pmatrix}.$$

Remark 7.3. Notice that in the neigborhood of the quadruple $\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ given in subsection 7.3 where the miniversal deformation is defined, the stratification induced (the strata are the intersection of the strata in $\mathcal{M}_{n,m,p}$ with the neigborhood) is Whitney regular. Then the transversal families defined over this neigborhood form an open and dense set.

Acknowledgments. The author is pleased to thank the referees for their valuable suggestions.

REFERENCES

[1] V.I. Arnold, *On matrices depending on parameters*, Uspekhi Mat. Nauk, 26 (1971), pp. 101–114.
[2] J. Berg and H. Kwatny, *A canonical parametrization of the Kronecker form of a matrix pencil*, Automatica, 31 (1995), pp. 669–680.
[3] J. Demmel and A. Edelman, *The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms*, Linear Algebra Appl., 230 (1995), pp. 61–88.
[4] A. Edelman, E. Elmroth, and B. Kågström, *A Geometric Approach to Perturbation Theory of Matrices and Matrix Pencils. Part I: Versal Deformation*, Tech. Report UMINF-93.22, Dept. of Computer Science, Umeå University, Umeå, Sweden, 1995.
[5] J. Ferrer and Mª I. García-Planas, *Structural stability of quadruples of matrices*, Linear Algebra Appl., 241 (1996), pp. 241–243.
[6] Mª I. García-Planas, *Estudio Geométrico de Familias Diferenciables de Parejas de Matrices*, Ph.D. thesis, Universitat Politècnica de Catalunya, Spain, 1994.
[7] C.G. Gibson, *Regularity of the Segre stratification,* Math. Proc. Cambridge Philos. Soc., 80 (1976), pp. 91–97.
[8] C.G. Gibson, K. Wirthmüller, A.A. du Plessis, and E.J.N. Looijenga, *Topological Stability of Smooth Mappings,* Springer-Verlag, New York, 1976.
[9] J. E. Humphreys, *Linear Algebraic Groups,* Springer-Verlag, New York, 1975.
[10] B.P. Molinari, *Structural invariants of linear multivariable systems*, Internat. J. Control, 28 (1978), pp. 493–510.
[11] A.S. Morse, *Structural invariants of linear multivariable systems*, SIAM J. Control, 11 (1973), pp. 446–465.
[12] A. Tannenbaum, *Invariance and System Theory: Algebraic and Geometric Aspects*, Lecture Notes in Math. 845, Springer-Verlag, New York, 1981.
[13] K. Tchon, *Bifurcation diagrams for feedback families of linear system*, Systems Control Lett., 5 (1985), pp. 397–401.
[14] J.S. Thorp, *The singular pencil of a linear dynamical system*, Internat. J. Control, 18 (1973), pp. 577–596.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 886–897, October 1998

004

EIGENVALUE LOCATIONS OF GENERALIZED COMPANION PREDICTOR MATRICES*

LICIO H. BEZERRA† AND FERMIN S. V. BAZÁN†

Abstract. Generalized predictor companion matrices arise in the linear prediction approach for the fit of a weighted sum of n exponentials to a given set of data points. They are special solutions of matrix equations of the type $\mathsf{H}(l+p)\,\mathsf{S} = \mathsf{H}(l)$, where for each $l \geq 0$ $\mathsf{H}(l)$ is an $M \times N$ Hankel matrix obtained from this data ($M \geq N > n$). We discuss in this paper results about the eigenvalue locations of this class of solutions by means of linear algebra techniques. An application of these results in the case that all the exponents have either negative or positive real parts is that the n exponentials can correspond to eigenvalues which are outside the unit circle depending on the choice of generalized predictor companion matrices. The other $(N-n)$ eigenvalues of these matrices always lie inside the unit circle and approach zero when p increases. This separation can facilitate their numerical calculation.

Key words. companion matrices, eigenvalues, linear prediction, exponential approximation

AMS subject classifications. 15A18, 65F15

PII. S0895479897314930

1. Introduction.

The identification of the parameters of functions

$$h(t) = r_1 e^{s_1 t} + \cdots + r_n e^{s_n t}, \quad \operatorname{Re} s_i < 0, \quad i = 1, ..., n, \tag{1.1}$$

is a problem which has been studied by several researchers in a variety of disciplines such as signal processing, mechanical vibrations, harmonic retrieval, acoustics, nuclear magnetic resonance, etc. One of the approaches to the problem is the linear prediction technique, which describes the problem as a matrix equation of the type $\mathsf{H}(l+1) = \mathsf{H}(l)\,\mathsf{S}$, where $\mathsf{H}(l)$ is an $M \times N$ Hankel matrix whose i, j entries are $h_{l+i+j-2} = h((l+i+j-2)\Delta t)$ with Δt as the sampling interval [11], [12]. This equation will be called a *prediction equation* of the system if both M and N are greater than or equal to n. Observe that, if the available data is free of noise, n is the rank of $\mathsf{H}(l)$ for any l (with respect to numerical rank determination; see, e.g., [2], [6], [17]). Usually we choose $M \geq N > n$ so that $\mathsf{H}(l)$ is rank deficient, and thus there are an infinite number of solutions of the prediction equation, which are called *predictor matrices* of the system. An important result is that n of the eigenvalues of S are $e^{s_1 \Delta t}, ..., e^{s_n \Delta t}$, which are called *system eigenvalues*. The parameters $r_1, ..., r_n$ are calculated once these eigenvalues are known.

In practical parameter identification problems the Hankel data matrices are corrupted by noise, which is assumed to be additive: $\tilde{\mathsf{H}}(l) = \mathsf{H}(l) + \mathcal{E}_l$. Several algorithms have been proposed for solving these problems, and an analysis of some available algorithms is presented in [19]. The prediction methods are based on solutions of

$$\tilde{\mathsf{H}}(l)x = \tilde{\mathsf{H}}(l+1)e_N, \tag{1.2}$$

*Received by the editors January 8, 1997; accepted for publication (in revised form) by S. Van Huffel October 2, 1997; published electronically May 27, 1998. Part of this research was performed while the first author was at CERFACS, Toulouse, France, and was supported by Fundação CAPES, Brasil, grant BEX 3119/95-13. The research of the second author was supported by CNPq, Brasil, grant 300487/94-0 (NV).

http://www.siam.org/journals/simax/19-4/31493.html

†Departamento de Matemática, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina 88040-900, Brasil (licio@mtm.ufsc.br, fermin@mtm.ufsc.br).

where e_N is the transpose of the canonical vector $(0 \cdots 01)$, for example, the Minimum-Norm method [9], in which x is the minimum norm least squares solution. The different possible choices for the dimensions M, N of $\tilde{\mathsf{H}}(l)$ and the ways of solving (1.2) have originated several estimation methods. For example, in [15] the Total Least Squares (TLS) approach is applied for solving the equation (1.2) to diminish the noise effects from the data matrices. Other methods are the Single Shift-Invariant and the Subspace Fitting methods (see again [19]), which include several algorithms like ESPRIT [14], HTLS [20], Kung's method [10], MUSIC [16], etc. Contributions for parameter estimation problems also appear in [1], [4], [7], [11], [12], and [21], among others. However, we don't intend to introduce here a better performing method for identifying parameters in the practical sense. The goal of this paper is to present theoretical results about the locations of the eigenvalues of a class of solutions of the general prediction equation $\mathsf{H}(l+p) = \mathsf{H}(l)\,\mathsf{S}$, $p \neq 0$, referred to here as generalized companion predictor matrices. These matrices naturally arise from the concept of linear prediction. Since any solution S is an $N \times N$ matrix and $N > n$, S has *extraneous* eigenvalues in addition to the system eigenvalues. Therefore, one problem is how to distinguish the system eigenvalues from the entire spectra of S, supposing the data free of noise. Our goal is to show that if the n exponentials are such that their exponents have either negative or positive real parts, then there are solutions S for which the system eigenvalues have absolute value greater than 1 while the extraneous ones have modulii less than 1. We begin with $p = 1$ and show that when S is the companion matrix $\mathcal{C}(c) = [e_2\, e_3 \cdots e_N\, c]$, where e_i is the ith canonical column vector and c is the minimum 2-norm least squares solution of $\mathsf{H}(l)\,x = \mathsf{H}(l+1)\,e_N$, then the extraneous eigenvalues are located inside the unit circle. Observe that the system eigenvalues corresponding to undamped signals ($\operatorname{Re} s_i > 0$) should lie outside the unit circle. It is worth emphasizing that the eigenvalue locations of companion predictor matrices was studied earlier, for example, by Kumaresan [9], in the context of the analysis of zeros of linear prediction-error filter polynomials for a class of deterministic signals. Here, we state results about these locations in a linear algebra context. This is done in section 2. Also in section 2, we show that the companion matrix $\hat{\mathcal{C}}(b) = [b\, e_1\, e_2 \cdots e_{N-1}]$, which is the solution of the backward prediction equation $\mathsf{H}(l-1) = \mathsf{H}(l)\,\mathsf{S}$, with b being the minimum 2-norm least squares solution of $\mathsf{H}(l)\,x = \mathsf{H}(l-1)\,e_1$, has the extraneous eigenvalues equal to the conjugates of the extraneous eigenvalues of $\mathcal{C}(c) = [e_2\, e_3 \cdots e_N\, c]$. Now, the corresponding system eigenvalues, $e^{-s_1\Delta t}, ..., e^{-s_n\Delta t}$, lie outside the unit circle if all the signals are damped ($\operatorname{Re} s_i < 0$). Results about eigenvalues of companion matrices can also be found in [3], [5], [8], [18]. In section 3, we introduce the concept of generalized companion predictor matrices as a class of solutions of the equation $\mathsf{H}(l+p) = \mathsf{H}(l)\,\mathsf{S}$, $p \neq 0, 1, -1$. We also present some results about the locations of the eigenvalues of these matrices. We finish this paper with numerical examples and some remarks.

2. Companion predictor matrices. Let $\mathsf{H}(l)$ be an $M \times N$ Hankel matrix whose entries are samples of $h(t)$:

$$\mathsf{H}(l) = [\vec{\mathbf{h}}_l \;\; \vec{\mathbf{h}}_{l+1} \;\; \cdots \;\; \vec{\mathbf{h}}_{l+N-1}] = \begin{bmatrix} h_l & h_{l+1} & \cdots & h_{l+N-1} \\ h_{l+1} & h_{l+2} & \cdots & h_{l+N} \\ \vdots & \vdots & & \vdots \\ h_{l+M-1} & h_{l+M} & \cdots & h_{l+M+N-2} \end{bmatrix}. \tag{2.1}$$

Then by (1.1), for all $l \geq 0$,

$$\mathsf{H}(l) = \mathsf{V}\Lambda^l \mathsf{R}\mathsf{W}^T, \tag{2.2}$$

where $\mathsf{V} = \mathsf{V}(\lambda_1, ..., \lambda_n)$ is the $M \times n$ Vandermonde matrix described by

$$\mathsf{V} = \begin{bmatrix} 1 & \cdots & 1 \\ \lambda_1 & \cdots & \lambda_n \\ \vdots & & \vdots \\ \lambda_1^{M-1} & \cdots & \lambda_n^{M-1} \end{bmatrix}, \tag{2.3}$$

$\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$, with $\lambda_j = e^{s_j \Delta t}$, $\mathsf{R} = \mathrm{diag}(r_1, ..., r_n)$, and W is the submatrix of V formed by taking its N first rows. A direct consequence of this decomposition is that, for all $l \geq 0$, $\mathrm{rank}(\mathsf{H}(l)) = n$ whenever $M \geq N \geq n$ and $\lambda_i \neq \lambda_j$ for $i \neq j$.

A predictor matrix, that is, a matrix S such that $\mathsf{H}(l+1) = \mathsf{H}(l)\mathsf{S}$, gives the new data sample $\vec{h}_{l+N}$ from the preceding N samples $\vec{h}_l, \vec{h}_{l+1}, ..., \vec{h}_{l+N-1}$. Observe that there can be an infinite number of matrices S satisfying this equation and that the parameters λ_i can be found from the eigenvalues of any predictor matrix according to the following relation:

$$\mathsf{H}(l+1) = \mathsf{H}(l)\mathsf{S}$$

$$\Updownarrow$$

$$\mathsf{W}^T\mathsf{S} = \Lambda\mathsf{W}^T. \tag{2.4}$$

LEMMA 2.1. *Let* S *be a solution of* (2.4), *where* W *is any* $N \times n$ *matrix and* $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$. *Then* $\mathcal{N}(\mathsf{W}^T)$ *is an invariant subspace under* S. *Furthermore, if* $\lambda_i \neq 0$ *for all* i, $\mathsf{S}\mathcal{N}(\mathsf{W}^T) = \mathcal{N}(\mathsf{W}^T)$.

Proof. $x \in \mathcal{N}(\mathsf{W}^T) \Leftrightarrow \mathsf{W}^T x = 0 \Rightarrow 0 = \Lambda\mathsf{W}^T x = \mathsf{W}^T\mathsf{S}x \Leftrightarrow \mathsf{S}x \in \mathcal{N}(\mathsf{W}^T)$ (if $\lambda_i \neq 0$ for all i, $\Rightarrow$ can be replaced by $\Leftrightarrow$ in the above chain). □

LEMMA 2.2. *Let* S *be a solution of* (2.4), *where* W *is a full rank* $N \times n$ *matrix* ($n \leq N$) *and* $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$. *Let* P *be a* $N \times (N-n)$ *matrix whose columns are an orthonormal basis of* $\mathcal{N}(\mathsf{W}^T)$, *and let* Q *be an* $N \times n$ *matrix whose columns are an orthonormal basis of the row space of* W^H, *that is,* $\mathcal{N}(\mathsf{W}^T)^\perp$. *Then, for* $x \neq 0$, $\mathsf{P}^H\mathsf{S}\mathsf{P}x = \lambda x$ *if and only if* $\mathsf{S}\mathsf{P}x = \lambda\mathsf{P}x$. *Moreover,* $\lambda(\mathsf{S}) = \lambda(\mathsf{Q}^H\mathsf{S}\mathsf{Q}) \cup \lambda(\mathsf{P}^H\mathsf{S}\mathsf{P}) = \lambda(\Lambda) \cup \lambda(\mathsf{P}^H\mathsf{S}\mathsf{P})$.

Proof. From Lemma 2.1, the columns of $\mathsf{S}\mathsf{P}$ form a basis for $\mathcal{N}(\mathsf{W}^T)$. Then, since P is a full rank matrix, given any vector x there is a unique vector y such that $\mathsf{P}y = \mathsf{S}\mathsf{P}x$: $y = \mathsf{P}^H\mathsf{S}\mathsf{P}x$. Therefore, $\mathsf{P}^H\mathsf{S}\mathsf{P}x = \lambda x \Leftrightarrow \mathsf{S}\mathsf{P}x = \lambda\mathsf{P}x$. Now, observe that

$$\begin{pmatrix} \mathsf{Q}^H \\ \mathsf{P}^H \end{pmatrix} \mathsf{S} \begin{pmatrix} \mathsf{Q} & \mathsf{P} \end{pmatrix} = \begin{pmatrix} \mathsf{Q}^H\mathsf{S}\mathsf{Q} & 0 \\ \mathsf{P}^H\mathsf{S}\mathsf{Q} & \mathsf{P}^H\mathsf{S}\mathsf{P} \end{pmatrix}.$$

Moreover, as $\overline{\mathsf{W}} = \mathsf{Q}\mathsf{Z}$, for some $n \times n$ nonsingular matrix Z, then

$$\mathsf{Q}^H\mathsf{S}\mathsf{Q} = \mathsf{Z}^{-H}\mathsf{Z}^H\mathsf{Q}^H\mathsf{S}\mathsf{Q} = \mathsf{Z}^{-H}\mathsf{W}^T\mathsf{S}\mathsf{Q} =$$

$$= \mathsf{Z}^{-H}\Lambda\mathsf{W}^T\mathsf{Q} = \mathsf{Z}^{-H}\Lambda\mathsf{Z}^H\mathsf{Q}^H\mathsf{Q} = \mathsf{Z}^{-H}\Lambda\mathsf{Z}^H. \qquad \square$$

Remark 2.3. By equation (2.2), the rows of $\mathsf{H}(l)$ for any l are spanned by the rows of W^T. So, the matrix Q in Lemma 2.2 can be calculated from a QR decomposition

of $\mathsf{H}(l)^H$. By Lemma 2.2, the characteristic polynomial of S, $p(x)$, can be written as $p(x) = (x-\lambda_1)\cdots(x-\lambda_n)g(x)$. Since one of our goals is to identify the λ_i, $i = 1, ..., n$, then it is important to know the properties of the $N-n$ roots of $g(x)$, which will be called *extraneous roots* from now on. Observe that $g(x)$ is the characteristic polynomial of $\mathsf{P}^H\mathsf{SP}$. We will focus our attention on a special solution S of equation (2.4)—the companion predictor matrix:

$$\mathcal{C} = [e_2\ e_3\ \cdots\ e_N\ c] = \begin{bmatrix} 0 & 0 & \cdots & 0 & c_0 \\ 1 & 0 & \cdots & 0 & c_1 \\ 0 & 1 & \cdots & 0 & c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & c_{N-1} \end{bmatrix}_{N\times N} \tag{2.5}$$

in which the column vector c is the minimum 2-norm solution of the system

$$\mathsf{H}(l)\, c = \mathsf{H}(l+1)e_N, \tag{2.6}$$

or, equivalently, of the following system:

$$\mathsf{W}^T c = \Lambda \mathsf{W}^T e_N. \tag{2.7}$$

PROPOSITION 2.4. *Let $\mathsf{W} = \mathsf{W}(\lambda_1, ..., \lambda_n)$ be a Vandermonde matrix of order $N \times n$, $N > n$, where $\lambda_i \neq \lambda_j$ for $i \neq j$, and let $\mathcal{C} = \mathcal{C}(c)$ be a companion matrix whose Nth column vector, $c = (c_0, ..., c_{N-1})$, is a solution of (2.7). If $\mu_1, ..., \mu_{N-n}$ are the extraneous roots of the characteristic polynomial of $\mathcal{C}$, then an eigenvector associated with μ_i is the vector of coefficients of the polynomial $(x-\lambda_1)\cdots(x-\lambda_n)(x-\mu_1)\cdots(x-\mu_{i-1})(x-\mu_{i+1})\cdots(x-\mu_{N-n})$.*

Proof. Since $\mathcal{C}$ is a solution of (2.4), an eigenvector associated with an extraneous root belongs to $\mathcal{N}(\mathsf{W}^T)$. Without loss of generality we may consider only the case of μ_1. So, let $a = (a_0, ..., a_{N-1})$ be an eigenvector associated with μ_1. Then $\mathcal{C}a = \mu_1 a$ is equivalent to

$$\begin{cases} c_0 a_{N-1} & = & \mu_1 a_0, \\ a_0 + c_1 a_{N-1} & = & \mu_1 a_1, \\ \vdots & \vdots & \vdots \\ a_{N-2} + c_{N-1} a_{N-1} & = & \mu_1 a_{N-1}. \end{cases}$$

Since $a_{N-1} \neq 0$ (otherwise $a = 0$), we can write

$$c_0 = \frac{\mu_1 a_0}{a_{N-1}},\ c_1 = \frac{\mu_1 a_1 - a_0}{a_{N-1}}, ..., c_{N-1} = \frac{\mu_1 a_{N-1} - a_{N-2}}{a_{N-1}}.$$

On the other hand, since $a \in \mathcal{N}(\mathsf{W}^T)$,

$$a(x) = a_{N-1}x^{N-1} + \cdots + a_1 x + a_0 = a_{N-1}(x-\lambda_1)\cdots(x-\lambda_n)(x-k_1)\cdots(x-k_{N-n-1}).$$

By comparing the coefficients of $a(x)$ to the coefficients of $c(x) = x^N - (c_{N-1}x^{N-1} + \cdots + c_1 x + c_0)$, we conclude that $c(x) = \frac{1}{a_{N-1}}\, a(x)(x-\mu_1)$. □

In order to prove the first proposition about the location of the eigenvalues of this matrix, we first state the following lemma, which is easily verified.

LEMMA 2.5. *Let P be an $m \times n$ matrix whose columns are orthonormal. If B is an $r \times s$ submatrix of P, then $||B||_2 \leq 1$.*

PROPOSITION 2.6. *Let* $\mathsf{W} = \mathsf{W}(\lambda_1, ..., \lambda_n)$ *be a Vandermonde matrix of order* $N \times n$, $N > n$, *where* $\lambda_i \neq \lambda_j$ *for* $i \neq j$, *and let* $\mathcal{C} = \mathcal{C}(c)$ *be a companion matrix whose* N*th column vector,* $c = (c_0, ..., c_{N-1})$, *is the minimum* 2*-norm solution of* (2.7). *Then the* $N - n$ *extraneous roots of the characteristic polynomial of* $\mathcal{C}$ *have moduli less than* 1.

Proof. Since $\mathcal{C}$ is a solution of (2.4), by Lemma 2.2 the extraneous roots are the ones of the characteristic polynomial of $\mathsf{P}^H\mathcal{C}\mathsf{P}$, where P is such that its columns form an orthonormal basis of $\mathcal{N}(\mathsf{W}^T)$. Since c is the minimum norm solution of a linear system, whose matrix is W^T, $c \in \mathcal{N}(\mathsf{W}^T)^\perp$. Hence, $\mathsf{P}^H\mathcal{C} = \begin{pmatrix} B^H & 0 \end{pmatrix}$, where B is a submatrix of P. Let μ be an extraneous root. Now, by Lemma 2.2 the extraneous roots are the roots of the characteristic polynomial of $\mathsf{P}^H\mathcal{C}\mathsf{P}$, and an eigenvector y of $\mathcal{C}$ associated with an extraneous eigenvalue is of the form $y = \mathsf{P}x$ for some x. If $\mu \neq 0$ then y has its last coordinate a different from 0 (if not, $\mathcal{C}y = \mu y \Rightarrow y = 0$). So, if $y = \begin{pmatrix} v \\ a \end{pmatrix} = \mathsf{P}x$ is an eigenvector of $\mathcal{C}$ associated with μ such that $||y|| = 1$ (therefore $||x|| = 1$), then

$$|\mu| = |\mu x| = ||\mathsf{P}^H\mathcal{C}\mathsf{P}x|| = || \begin{pmatrix} B^H & 0 \end{pmatrix} \begin{pmatrix} v \\ a \end{pmatrix} || = ||B^H v|| \leq ||B^H|| ||v||.$$

By Lemma 2.5, $||B^H||_2 \, ||v||_2 \leq ||v||_2$. And in the case of 2-norm, $||v|| < ||\begin{pmatrix} v \\ a \end{pmatrix}|| = 1$. □

Remark 2.7. It can happen that a companion matrix as defined in Proposition 2.4 can have a system eigenvalue as an extraneous eigenvalue, as in the following example.

Example 2.8. Let $\lambda_1 = 1$ and λ_2 be the real root of the polynomial $2x^3 + 3x^2 + 4x + 1$. Let

$$\mathsf{W}^T = \begin{bmatrix} 1 & \lambda_1 & \lambda_1^2 \\ 1 & \lambda_2 & \lambda_2^2 \end{bmatrix}.$$

Let $\mathcal{C}(c)$ be the companion matrix such that $c = (c_0, c_1, c_2)$ is the minimum 2-norm solution of (2.7). Then $x^3 - c_2x^2 - c_1x - c_0 = (x - \lambda_1)(x - \lambda_2)^2$. Therefore, although λ_2 is the only extraneous root, its geometric multiplicity is 1 and not 2 because a companion matrix is nonderogatory [22], [13].

Since the parameters s_i in the function $h(t)$ have negative real parts, the nonextraneous eigenvalues $e^{s_i \Delta t}$ of any solution S of (2.4) have moduli less than 1. Our goal is to identify these eigenvalues among all the eigenvalues of S. The above proposition states that when $\mathsf{S} = \mathcal{C}(c)$ the extraneous eigenvalues also have moduli less than 1. At first sight this doesn't help us in the task of identification of the nonextraneous roots. However, when the solution $\hat{\mathcal{C}}(b)$ of the backward prediction equation

$$\mathsf{H}(l+1)\, b = \mathsf{H}(l) e_1 \tag{2.8}$$

is considered, the nonextraneous eigenvalues have moduli greater than 1, while the extraneous roots, which are the conjugates of the extraneous roots of $\mathcal{C}(c)$, have moduli less than 1. A separation between the two sets of eigenvalues is then realized. This companion matrix, which from now on will be called companion *b-predictor* matrix,

is

$$\hat{\mathcal{C}} = \hat{\mathcal{C}}(b) = [b\ e_1\ e_2\ \cdots\ e_{N-1}] = \begin{bmatrix} b_{N-1} & 1 & 0 & \cdots & 0 \\ b_{N-2} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_1 & 0 & 0 & \cdots & 1 \\ b_0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{N\times N}, \tag{2.9}$$

which satisfies

$$\mathsf{H}(l+1)\hat{\mathcal{C}} = \mathsf{H}(l) \Leftrightarrow \mathsf{W}^T\hat{\mathcal{C}} = \Lambda^{-1}\mathsf{W}^T. \tag{2.10}$$

As before, the column vectors of W are left eigenvalues of $\hat{\mathcal{C}}$, which are now related to $\lambda_1^{-1}, ..., \lambda_n^{-1}$. One can see that $\hat{\mathcal{C}}(b) = \mathcal{P}^T\mathcal{C}(\hat{b})\mathcal{P}$, where $\mathcal{P} = \mathcal{P}^T$ is the $N \times N$ permutation matrix such that $\mathcal{P}e_i = e_{N-i}$ and $\hat{b} = \mathcal{P}b$.

Remark 2.9. Another result obtained in a similar way to Proposition 2.4 is that if $\mu_1, ..., \mu_r$ are the nonzero extraneous roots of the characteristic polynomial of $\hat{\mathcal{C}}$, the companion matrix whose first column vector, $b = (b_{N-1}, ..., b_0)$, is a solution of (2.10), then an eigenvector associated with μ_i is the vector of coefficients of the polynomial $(x-\lambda_1)\cdots(x-\lambda_n)(x-\mu_1^{-1})\cdots(x-\mu_{i-1}^{-1})(x-\mu_{i+1}^{-1})\cdots(x-\mu_r^{-1})$. This means that if we have an extraneous root, then an eigenvector associated with it gives us the other extraneous roots in addition to the system eigenvalues.

PROPOSITION 2.10. *Let* $\mathsf{W} = \mathsf{W}(\lambda_1, ..., \lambda_n)$ *be a Vandermonde matrix of order* $N \times n$, $N > n$, *where* $\lambda_i \neq \lambda_j$ *for* $i \neq j$, $\lambda_i \neq 0$ *for all* i. *Let* $\hat{\mathcal{C}} = \hat{\mathcal{C}}(b)$ *be a companion* b*-predictor matrix whose first column vector,* $b = (b_{N-1}, ..., b_0)$, *is the minimum* 2*-norm solution of* (2.8). *Then the nonextraneous eigenvalues of* $\hat{\mathcal{C}}$ *are* λ_i^{-1}, $i = 1, ..., n$, *and the* $(N-n)$ *extraneous roots are the conjugates of the* $(N-n)$ *extraneous roots of* $\mathcal{C} = \mathcal{C}(c)$, *the companion matrix where* c *is the minimum* 2*-norm solution of equation* (2.7).

Proof. By (2.10) the column vectors of W are left eigenvectors of $\hat{\mathcal{C}}$ corresponding to the eigenvalues $\lambda_1^{-1}, ..., \lambda_n^{-1}$.

In order to see that the $N-n$ extraneous eigenvalues of $\mathcal{C}$ are the conjugates of the $N-n$ extraneous eigenvalues of $\hat{\mathcal{C}}$, let $p(z)$ and $\hat{p}(z)$ be defined as

$$p(z) = z^N - c_{N-1}z^{N-1} - \cdots - c_1 z - c_0$$

and

$$\hat{p}(z) = z^N - b_{N-1}z^{N-1} - \cdots - b_1 z - b_0,$$

respectively. Then $p(z) = f(z)g(z)$ and $\hat{p}(z) = \hat{f}(z)\hat{g}(z)$, where $f(z) = (z-\lambda_1)\cdots(z-\lambda_n)$ and $\hat{f}(z) = (z-\lambda_1^{-1})\cdots(z-\lambda_n^{-1})$. The problem is then to show that the roots of $g(z)$ are the conjugates of the roots of $\hat{g}(z)$. Or equivalently, that the points which minimize the function

$$F = F(z_1, ..., z_{N-n}) = 1 + |F_1(z_1, ..., z_{N-n})|^2 + \cdots + |F_N(z_1, ..., z_{N-n})|^2$$

are the conjugates of the points which minimize

$$\hat{F} = \hat{F}(z_1, ..., z_{N-n}) = 1 + |\hat{F}_1(z_1, ..., z_{N-n})|^2 + \cdots + |\hat{F}_N(z_1, ..., z_{N-n})|^2,$$

where

$$z^N - F_1 z^{N-1} - \cdots - F_N = f(z)(z-z_1)\cdots(z-z_{N-n})$$

and

$$z^N - \hat{F}_1 z^{N-1} - \cdots - \hat{F}_N = \hat{f}(z)(z - z_1)\cdots(z - z_{N-n}).$$

On the other hand,

$$2\pi i\, F = \int_{\mathcal{B}} \bar{z} f(z)\overline{f(z)}\, |z - z_1|^2 \cdots |z - z_{N-n}|^2 dz,$$

where $\mathcal{B}$ is the unit circumference. Moreover, for $z \in \mathcal{B}$, $z = \bar{z}^{-1}$, and so,

$$f(z)\overline{f(z)} = |\lambda_1|^2 \cdots |\lambda_n|^2 \hat{f}(\bar{z})\overline{\hat{f}(\bar{z})}.$$

But $\hat{f}(\bar{z})\overline{\hat{f}(\bar{z})} = |\bar{z} - \lambda_1^{-1}|^2 \cdots |\bar{z} - \lambda_n^{-1}|^2 = |z - \overline{\lambda_1^{-1}}|^2 \cdots |z - \overline{\lambda_n^{-1}}|^2$. Let $d(z) = (z - \lambda_1^{-1}) \cdots (z - \overline{\lambda_n^{-1}})$. So,

$$F = |\lambda_1|^2 \cdots |\lambda_n|^2 D, \text{ where } 2\pi i\, D = \int_{\mathcal{B}} \bar{z} d(z)\overline{d(z)}\, |z - z_1|^2 \cdots |z - z_{N-n}|^2 dz.$$

Therefore, F and D have the same minimum points. But the minimum points of D are the conjugates of the ones of $\hat{F}$. □

Remark 2.11. Since the eigenvalues of a matrix depend continuously on its entries, given a companion b-predictor matrix $\hat{\mathcal{C}}(b)$, where b is the minimum 2-norm solution of the prediction system, there is a neighborhood $\mathcal{V}$ of b in $\mathbb{C}^N$ such that the companion matrices $\hat{\mathcal{C}}(\hat{b})$ still have n eigenvalues with moduli greater than 1 and $(N - n)$ eigenvalues with moduli less than 1 for all $\hat{b} \in \mathcal{V}$.

3. Generalized companion predictor matrices. We shall now generalize the notion of a companion predictor matrix. Since for all $l \geq 0$ we have $\mathsf{H}(l+1) = \mathsf{H}(l)\mathcal{C}$, then for any positive integer p, $\mathsf{H}(l+p) = \mathsf{H}(l+p-1)\mathcal{C} = \mathsf{H}(l+p-2)\mathcal{C}^2 = \cdots = \mathsf{H}(l)\mathcal{C}^p$. So, it is possible to compute the state $l + p$ directly from the state l. This motivates the following definition.

DEFINITION 3.1. *Let* $\mathsf{H}(l)$ *be a Hankel matrix defined as in* (2.1). *Let* p *be a positive integer.* S_p *is a* p*-predictor matrix if, for all* $l \geq 0$, $\mathsf{H}(l + p) = \mathsf{H}(l)\mathsf{S}_p$.

From the above definition and (2.2), we have

$$\mathsf{H}(l + p) = \mathsf{H}(l)\mathsf{S}_p \Leftrightarrow \mathsf{W}^T\mathsf{S}_p = \Lambda^p \mathsf{W}^T. \tag{3.1}$$

There is a collection of matrices which satisfy the above definition, and for all these matrices S_p, W^T is a matrix of left eigenvectors of S_p associated with $\lambda_1^p, ..., \lambda_n^p$. We shall analyze the eigenvalue location of a class of p-predictor matrices, $1 \leq p < N$, which in some sense are a sort of generalized companion matrices:

$$\mathcal{C}_p = [e_{p+1} \ \cdots \ e_N \ c^{(1)} \ \cdots \ c^{(p)}]. \tag{3.2}$$

Here, $c^{(i)}$, $i = 1, ..., p$, are column vectors satisfying the following system:

$$\mathsf{H}(l)c^{(i)} = \mathsf{H}(l + p)e_{N-i+1} \tag{3.3}$$

or, equivalently,

$$\mathsf{W}^T c^{(i)} = \Lambda^i \mathsf{W}^T e_N, \ i = 1, \ldots, p. \tag{3.4}$$

PROPOSITION 3.2. *Let $\mathcal{C}_p$ be a p-predictor matrix defined as in* (3.2) *with the column vectors $c^{(i)}$, $1 \le i \le p$, being the minimum 2-norm solutions of the system* (3.4), *where $\mathsf{W} = \mathsf{W}(\lambda_1, \ldots, \lambda_n)$ is an $N \times n$ Vandermonde matrix with $\lambda_i \neq \lambda_j$ if $i \neq j$, and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Then $\lambda_1^p, ..., \lambda_n^p$ are n of its eigenvalues and the extraneous roots have moduli less than* 1. *Moreover, if $1 \le p \le n$, then* $\operatorname{rank}(\mathcal{C}_p) \le n$; *otherwise,* $\operatorname{rank}(\mathcal{C}_p) \le N + n - p$.

Proof. Since $\mathcal{C}_p$ is also a solution of the equation $\mathsf{H}(l+p) = \mathsf{H}(l)\mathsf{S}_p$, we have by (3.1) that

$$\mathsf{W}^T \mathcal{C}_p = \Lambda^p \mathsf{W}^T. \tag{3.5}$$

Therefore, $\lambda_1^p, ..., \lambda_n^p$ are eigenvalues of $\mathcal{C}_p$. The proof that the extraneous roots have moduli less than 1 is analogous to the proof of Proposition 2.6, remarking that now the last p coordinates of the eigenvectors of $\mathcal{C}_p$ cannot be all null at the same time.

If $1 \le p \le n$, the set of vectors $\Lambda^i \mathsf{W}^T e_n$, $i = 1, ..., p$, is linearly independent. Let $(\mathsf{W}^T)^\dagger$ be the pseudoinverse of W^T. Then, since $(\mathsf{W}^T)^\dagger$ is a full rank matrix, the vectors $c^{(i)} = (\mathsf{W}^T)^\dagger \Lambda^i \ \mathsf{W}^T e_N$, $i = 1, ..., p$, are linearly independent. So, $p \le \operatorname{rank}(\mathcal{C}_p) \le n$. For $n < p \le N$, since for all i, $1 \le i \le p$, $c^{(i)} \in \mathcal{N}(\mathsf{W}^T)^\perp$, and $\dim \mathcal{N}(\mathsf{W}^T)^\perp = n$, $\operatorname{rank}(\mathcal{C}_p) \le N + n - p$. □

When linear prediction is carried out in the backward direction we have the following definition.

DEFINITION 3.3. *Let $\mathsf{H}(l)$ be a Hankel matrix defined as in* (2.1). *S_p^b is said to be a p-backward predictor matrix if, for $p > 0$ and $l \ge 0$, $\mathsf{H}(l) = \mathsf{H}(l+p)\mathsf{S}_p^b$.*

Analogous to the case of forward prediction, if $(\forall i)\ \lambda_i \neq 0$ and $\lambda_i \neq \lambda_j$ for $i \neq j$, then

$$\mathsf{H}(l) = \mathsf{H}(l+p)\mathcal{C}_p^b \Leftrightarrow \mathsf{W}^T \mathsf{S}_p^b = \Lambda^{-p} \mathsf{W}^T. \tag{3.6}$$

Hence, $\lambda_1^{-p}, ..., \lambda_n^{-p}$ belong to the spectrum of S_p^b. Again, we are interested in analyzing p-backward predictor matrices of the following type:

$$\mathcal{C}_p^b = [b^{(p)} \ \cdots \ b^{(1)} \ e_1 \ \cdots \ e_{N-p}], \tag{3.7}$$

where b_i is the minimum norm solution of

$$\mathsf{H}(l+p) b^{(i)} = \mathsf{H}(l) e_i, \ i = 1, ..., p, \tag{3.8}$$

which is equivalent to the following equation:

$$\mathsf{W}^T b^{(i)} = \Lambda^{-i} \mathsf{W}^T e_1, \ i = 1, ..., p. \tag{3.9}$$

PROPOSITION 3.4. *Let $\mathsf{W} = \mathsf{W}(\lambda_1, \ldots, \lambda_n)$ be a Vandermonde matrix of order $N \times n$, $N > n$, where $\lambda_i \neq \lambda_j$ for $i \neq j$, $\lambda_i \neq 0$ for all i. Let $p \le N$ and $\mathcal{C}_p^b = [b^{(1)} \ \cdots \ b^{(p)} \ e_1 \ \cdots \ e_{N-p}]$, where $b^{(i)}$ is the minimum* 2*-norm solution of* (3.9). *Then, $\lambda_1^{-p}, ..., \lambda_n^{-p}$ are n of its eigenvalues, whereas the extraneous roots are the conjugates of those of its corresponding matrix $\mathcal{C}_p$ (equation* (3.2)*).*

Proof. As $\mathcal{C}_p^b$ is also a solution of the equation $\mathsf{H}(l+p)\mathsf{S}_p^b = \mathsf{H}(l)$, we have by (3.6) that

$$\mathsf{W}^T \mathcal{C}_p^b = \Lambda^{-p} \mathsf{W}^T. \tag{3.10}$$

Therefore, $\lambda_1^{-p}, ..., \lambda_n^{-p}$ are eigenvalues of $\mathcal{C}_p^b$.

$\mathcal{N}(\mathsf{W}^T)$ is an invariant right subspace of both $\mathcal{C}_p$ and $\mathcal{C}_p^b$ (Lemma 2.1), which is associated with the extraneous roots (Lemma 2.2). Now, let $\mathcal{U} = (u_i^j)$ be a matrix whose $(N-n)$ columns form an orthonormal basis of $\mathcal{N}(\mathsf{W}^T)$. We wish to prove that $\mathcal{U}^H \mathcal{C}_p \mathcal{U}$ is equal to the conjugate of $\mathcal{U}^H \mathcal{C}_p^b \mathcal{U}$, which yields the statement about the extraneous roots. Since the last p columns of $\mathcal{C}_p$ as well as the first p columns of $\mathcal{C}_p^b$ are in $\mathcal{N}(\mathsf{W}^T)^\perp$, we have

$$\mathcal{U}^H \mathcal{C}_p \mathcal{U} = \begin{pmatrix} \bar{u}_{p+1}^1 & \cdots & \bar{u}_N^1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \bar{u}_{p+1}^{N-n} & \cdots & \bar{u}_N^{N-n} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} u_1^1 & \cdots & u_1^{N-n} \\ \vdots & & \vdots \\ u_{N-p}^1 & \cdots & u_{N-p}^{N-n} \\ u_{N-p+1}^1 & \cdots & u_{N-p+1}^{N-n} \\ \vdots & & \vdots \\ u_N^1 & \cdots & u_N^{N-n} \end{pmatrix}$$

and

$$\mathcal{U}^H \mathcal{C}_p^b \mathcal{U} = \begin{pmatrix} 0 & \cdots & 0 & \bar{u}_1^1 & \cdots & \bar{u}_{N-p}^1 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & \bar{u}_1^{N-n} & \cdots & \bar{u}_{N-p}^{N-n} \end{pmatrix} \begin{pmatrix} u_1^1 & \cdots & u_1^{N-n} \\ \vdots & & \vdots \\ u_p^1 & \cdots & u_p^{N-n} \\ u_{p+1}^1 & \cdots & u_{p+1}^{N-n} \\ \vdots & & \vdots \\ u_N^1 & \cdots & u_N^{N-n} \end{pmatrix}.$$

So,

$$\mathcal{U}^H \mathcal{C}_p^b \mathcal{U} = (\mathcal{U}^H \mathcal{C}_p \mathcal{U})^H. \qquad \square$$

Remark 3.5. Observe that the above demonstration provides another proof of Proposition 2.10.

Remark 3.6. When the eigenvalues λ_i are either real or complex conjugate pairs, both matrices $\mathcal{C}_p$ and $\mathcal{C}_p^b$ are real, and so, those invariant subspaces are also real.

Remark 3.7. If $\mathcal{U}$ is an orthonormal basis of $\mathcal{N}(\mathsf{W}^T)$, both $(\hat{\mathcal{C}})^p$ and $\mathcal{C}_p^b$ can be decomposed as

$$(\hat{\mathcal{C}})^p = (I - \mathcal{U}\mathcal{U}^H)\,(\hat{\mathcal{C}})^p + \mathcal{U}\mathcal{U}^H\,(\hat{\mathcal{C}})^p = \mathcal{L} + \mathcal{M}, \tag{3.11}$$

$$\mathcal{C}_p^b = (I - \mathcal{U}\mathcal{U}^H)\,\mathcal{C}_p^b + \mathcal{U}\mathcal{U}^H\,\mathcal{C}_p^b = \mathcal{L} + \hat{\mathcal{M}}, \tag{3.12}$$

where $\mathcal{U}\mathcal{U}^H$ is the orthogonal projection onto $\mathcal{N}(\mathsf{W}^T)$ and $I - \mathcal{U}\mathcal{U}^H$ is the orthogonal projection onto $\mathcal{N}(\mathsf{W}^T)^\perp$. Observe that $I - \mathcal{U}\mathcal{U}^H = (\mathsf{W}^T)^\dagger \mathsf{W}^T$ and $(\mathsf{W}^T)^\dagger \mathsf{W}^T\,(\hat{\mathcal{C}})^p = (\mathsf{W}^T)^\dagger \mathsf{W}^T\,\mathcal{C}_p^b = (\mathsf{W}^T)^\dagger \Delta^{-p}\,\mathsf{W}^T = \mathcal{L}$, because both matrices are solutions of (3.6). Hence, $(\hat{\mathcal{C}})^p - \mathcal{C}_p^b = \mathcal{M} - \hat{\mathcal{M}}$. Since the first p columns of $\mathcal{C}_p^b$ belong to $\mathcal{N}(\mathsf{W}^T)^\perp$, the first p columns of $\hat{\mathcal{M}} = \mathcal{U}\mathcal{U}^H\,\mathcal{C}_p^b$ are zero. Thus, $(\hat{\mathcal{C}})^p - \mathcal{C}_p^b = \mathcal{M} - \mathsf{N}^p$, where N is the nilpotent matrix such that $\mathsf{N}e_1 = 0$, $\mathsf{N}e_i = e_{i+1}$ for $i = 1, \ldots, N-1$. The nonzero extraneous roots of $\mathcal{C}_p^b$ are the nonzero eigenvalues of $\mathcal{X} = \mathcal{U}\mathcal{U}^H\mathcal{C}_p^b$, which equals $\mathcal{U}\mathcal{U}^H\mathsf{N}^p$. On the other hand, the nonzero eigenvalues of $\mathcal{X}$ are the nonzero eigenvalues of $\mathcal{U}\mathcal{U}^H_{(p+1:N,1:N-p)}$, that is, the submatrix of $\mathcal{U}\mathcal{U}^H$ with the rows and columns indicated by the subscript.

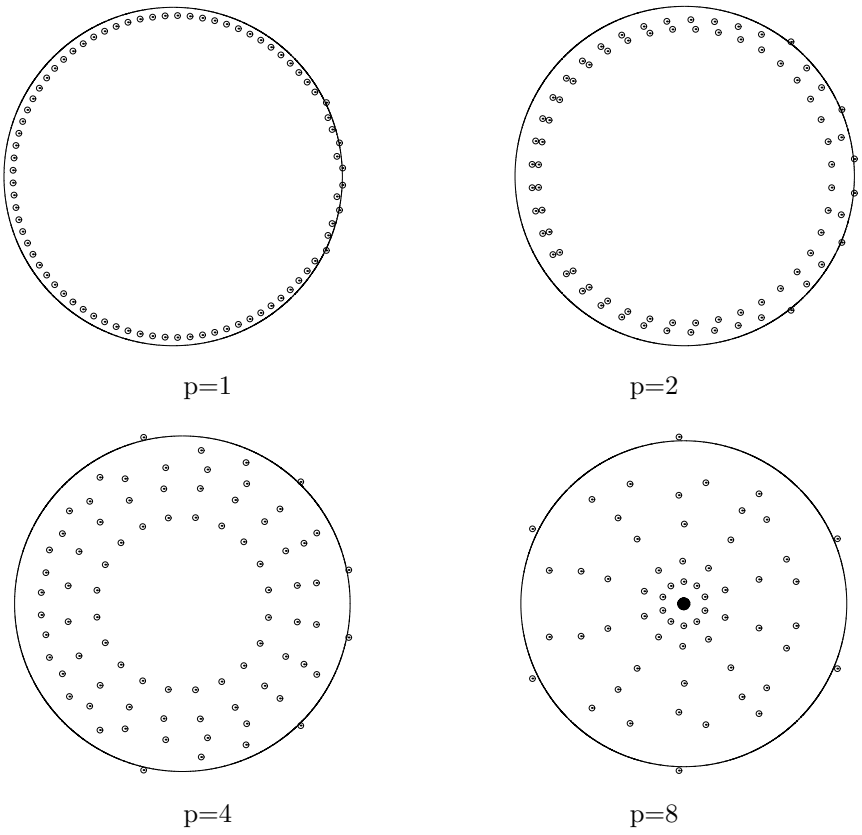


FIG. 3.1. *Eigenvalue locations of p-backward predictor matrices.*

We hope for a better separation of the system eigenvalues from the extraneous roots of p-backward predictor matrices when p increases, because $\mathcal{U}\mathcal{U}^H \mathsf{N}^p = 0$ when $p = N$. Thus, the nonzero extraneous roots approach zero when p increases. The system signals have moduli greater than 1 and they will increase in magnitude with p. Observe that if $|\mu_1| \geq |\mu_2| \geq \cdots \geq |\mu_{N-n}|$ are the extraneous roots of $\mathcal{C}_p^b$, then

$$\sum_{i=1}^{N-n} |\mu_i|^2 \leq ||\mathcal{U}\mathcal{U}^H \mathsf{N}^p||_F^2 = \sum_{i=1}^{N-p} ||\mathcal{U}\mathcal{U}^H \, e_i||_2^2.$$

Since $||\mathcal{C}_p^b||_F \leq ||(\hat{\mathcal{C}})^p||_F$, we also expect a better separation between the two classes of eigenvalues when the matrix $\mathcal{C}_p^b$ is used instead of $(\hat{\mathcal{C}})^p$.

In Fig. 3.1, we can see the behavior of the eigenvalues of $\mathcal{C}_p^b$, which is a p-backward predictor matrix for a three-dimensional representation of a simulated mechanical system whose impulse response function is

$$h(t) = e^{-0.06t} \sin(4t) + 0.8e^{-0.056t} \sin(t) + 1.2e^{-0.09t} \sin(9t),$$

taking $N = 80$ and $\Delta t = 0.05$. We observe in Fig. 3.1 that the locations of the extraneous roots seem to have a certain pattern. This is because these roots are eigenvalues of submatrices $\mathcal{P}_p$ corresponding to different p values of $\mathcal{P}_0 = \mathcal{U}\mathcal{U}^H$, with

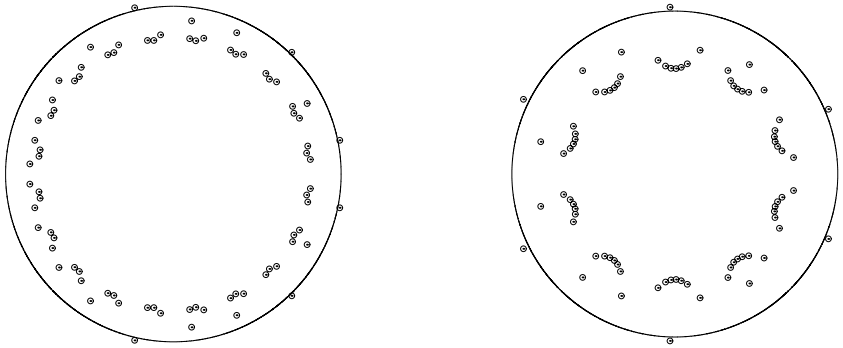


FIG. 3.2. *Eigenvalue locations of powers p of a 1-backward predictor matrix.*

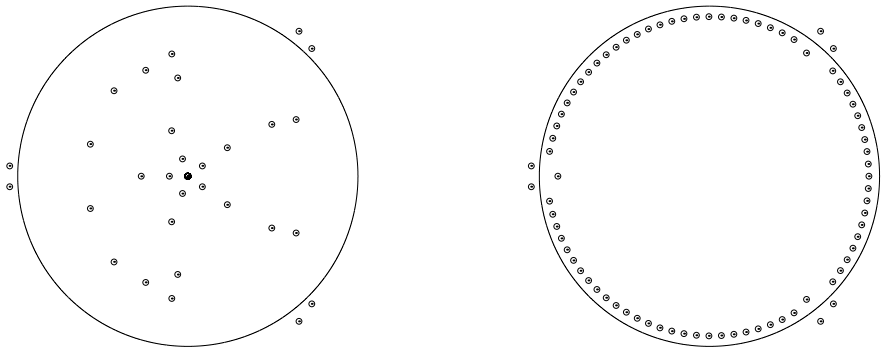


FIG. 3.3. *Eigenvalue locations of a 16-backward predictor matrix and a 1-backward predictor matrix after a downsampling operation.*

each submatrix $\mathcal{P}_r$ embedded in $\mathcal{P}_s$ if $r > s$. In Fig. 3.2 we can compare the locations of the extraneous roots of $(\hat{\mathcal{C}})^p$ with the corresponding roots of $\mathcal{C}_p^b$.

Concluding remarks. In this paper an eigenvalue locations analysis of generalized companion predictor matrices has been presented and illustrated by numerical examples. A final remark is that if the sample rate is reduced by a factor of p ($\widehat{\Delta t} = p\Delta t$), the resulting regular predictor matrix $\widehat{\mathcal{C}}^b$ has the same system eigenvalues as the generalized companion matrix $\mathcal{C}_p^b$ (sample rate equal to Δt). However, the extraneous eigenvalues are very different. This fact is illustrated for $p = 16$ in Fig. 3.3. The regular predictor matrix ($\widehat{\Delta t} = 0.8$) has their extraneous eigenvalues close to the unit circle, while the extraneous eigenvalues of $\mathcal{C}_{16}^b$ ($\Delta t = 0.05$) are closer to zero.

Acknowledgments. The authors wish to thank Alan McCoy (CERFACS) and Iain Duff (CERFACS/Rutherford Appleton Laboratory) for profitable discussions and helpful comments regarding the presentation of this paper.

REFERENCES

[1] F. S. V. Bazán and C. Bavastri, *An optimized pseudo-inverse algorithm (OPIA) for multi-input multi-output modal parameter identification*, Mechanical Systems and Signal Processing, 10 (1996), pp. 365–380.
[2] T. F. Chan, *Rank revealing QR factorizations*, Linear Algebra Appl., 88/89 (1987), pp. 67–82.
[3] G. Cybenko, *Locations of zeros of predictor polynomials*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 235–237.
[4] B. De Moor, *The singular value decomposition and long and short spaces of noisy matrices*, IEEE Trans. Signal Processing, 41 (1993), pp. 2826–2838.
[5] A. Edelman and H. Murakami, *Polynomial roots from companion matrix eigenvalues*, Math. Comp., 64 (1995), pp. 763–776.
[6] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.
[7] P. De Groen and B. De Moor, *The fit of a sum of exponentials to noisy data*, J. Comput. Appl. Math., 20 (1987), pp. 175–187.
[8] F. Kittaneh, *Singular values of companion matrices and bounds on zeros of polynomials*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 333–340.
[9] R. Kumaresan, *On the zeros of the linear prediction-error filters for deterministic signals*, IEEE Trans. on ASSP, ASSP-31 (1983), pp. 217–221.
[10] S. Y. Kung, K. S. Arun, and D. V. B. Rao, *State-space and singular value decomposition-based approximation methods for the harmonic retrieval problem*, J. Opt. Soc. Amer., 73 (1983), pp. 1799–1811.
[11] Z. Liang and D. J. Inman, *Matrix decomposition methods in experimental modal analysis*, J. Vibrations and Acoustics, 112 (1990), pp. 410–413.
[12] J. Makhoul, *Linear prediction: A tutorial review*, Proc. IEEE, 63 (1975), pp. 561–580.
[13] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*, Dover Publications, Mineola, NY, 1992.
[14] R. Roy and T. Kailath, ESPRIT-*estimation of signal parameters via rotational invariance techniques*, IEEE Trans. Acoust. Speech. Signal Process., ASSP-37 (1989), pp. 984–995.
[15] M. A. Rahman and K. B. Yu, *Total least square approach for frequency estimation using linear prediction*, IEEE Trans. Acoust. Speech Signal Process., ASSP-35 (1987), pp. 1440–1454.
[16] R. O. Schmidt, *Multiple emitter location and signal parameter estimation*, IEEE Trans. Antennas and Propagation, AP-34 (1986), pp. 276–281.
[17] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.
[18] K.-C. Toh and L. N. Trefethen, *Pseudozeros of polynomials and pseudospectra of companion matrices*, Numer. Math., 68 (1994), pp. 403–425.
[19] A. van der Veen, E. F. Deprettere, and A. L. Swindlehurst, *Subspace-based signal analysis using singular value decomposition*, Proc. IEEE, 81 (1993), pp. 1277–1309.
[20] S. Van Huffel, H. Chen, C. Decanniere, and P. Van Hecke, *Improved total least squares based algorithm for time-domain NMR data quantification*, J. Magn. Reson. A, 110 (1994), pp. 228–237.
[21] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, PA, 1991.
[22] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 898–905, October 1998

005

SOME RECENT RESULTS ON THE LINEAR COMPLEMENTARITY PROBLEM*

G. S. R. MURTHY†, T. PARTHASARATHY‡, AND B. SRIPARNA‡

Abstract. In this article we present some recent results on the linear complementarity problem. It is shown that (i) within the class of column adequate matrices, a matrix is in $\boldsymbol{Q_0}$ if and only if it is completely $\boldsymbol{Q_0}$ (ii) for the class of $\boldsymbol{C_0^f}$-matrices introduced by Murthy and Parthasarathy [*SIAM J. Matrix Anal. Appl.*, 16 (1995), pp. 1268–1286], we provide a sufficient condition under which a matrix is in $\boldsymbol{P_0}$ and as a corollary of this result, we give an alternative proof of the result that $\boldsymbol{C_0^f} \cap \boldsymbol{Q_0} \subseteq \boldsymbol{P_0}$ (iii) within the class of INS-matrices introduced by Stone [Department of Operations Research, Stanford University, Stanford, CA, 1981], a nondegenerate matrix must necessarily have the block property introduced by Murthy, Parthasarathy, and Sriparna [G. S. R. Murthy, T. Parthasarathy, and B. Sriparna, *Linear Algebra Appl.*, 252 (1997), pp. 323–337]. Furthermore, we conjecture that if a matrix has block property, then it must be Lipschitzian. This problem is an important one from two angles: if the conjecture is true, it provides a finite test to check whether a given matrix is Lipschitzian or nondegenerate *INS*; and it settles an open problem posed by Stone. It is shown that the conjecture is true in the cases of 2×2-matrices, nonnegative and nonpositive matrices of general order.

Key words. linear complementarity problem, adequacy, matrix classes, principal pivoting

AMS subject classification. 90C33

PII. S0895479896313814

1. Introduction.

Given a matrix $A \in \boldsymbol{R}^{n \times n}$ and $q \in \boldsymbol{R}^n$ the linear complementarity problem (LCP) is to find a vector $z \in \boldsymbol{R}^n$ such that

$$Az + q \geq 0, \quad z \geq 0, \text{ and } \quad z^t(Az + q) = 0. \tag{1.1}$$

LCP has numerous applications, both in theory and in practice, and is treated by a vast literature (see [2, 10]). Let $F(q, A) = \{ z \in \boldsymbol{R}^n_+ : Az + q \geq 0 \}$ and $S(q, A) = \{ z \in F(q, A) : (Az + q)^t z = 0 \}$. A number of matrix classes have been defined in connection with LCP, the fundamental ones being $\boldsymbol{Q}$ and $\boldsymbol{Q_0}$. The class $\boldsymbol{Q}$ consists of all real square matrices A such that $S(q, A) \neq \phi$ for every $q \in \boldsymbol{R}^n$ [11], and $\boldsymbol{Q_0}$ consists of all real square matrices A such that $S(q, A) \neq \phi$ whenever $F(q, A) \neq \phi$ [9].

For any positive integer n, write $\bar{n} = \{1, 2, \ldots, n\}$, and for any subset α of $\bar{n}$, write $\bar{\alpha} = \bar{n} \setminus \alpha$. For any $A \in \boldsymbol{R}^{n \times n}$, $A_{\alpha\alpha}$ is obtained by dropping rows and columns corresponding to $\bar{\alpha}$ from A. For any $x \in \boldsymbol{R}^n$, x_α is obtained from x by dropping coordinates corresponding to $\bar{\alpha}$, and x_i denotes the ith coordinate of x. Consider $A \in \boldsymbol{R}^{n \times n}$. If $\alpha \subseteq \bar{n}$ is such that $\det A_{\alpha\alpha} \neq 0$, then the matrix M defined by

$$M_{\alpha\alpha} = (A_{\alpha\alpha})^{-1}, \; M_{\alpha\bar{\alpha}} = -M_{\alpha\alpha}A_{\alpha\bar{\alpha}}, \; M_{\bar{\alpha}\alpha} = A_{\bar{\alpha}\alpha}M_{\alpha\alpha}, \; M_{\bar{\alpha}\bar{\alpha}} = A_{\bar{\alpha}\bar{\alpha}} - M_{\bar{\alpha}\alpha}A_{\alpha\bar{\alpha}}$$

is known as the principal pivotal transform (PPT) of A with respect to α and will be denoted by $\wp_\alpha(A)$. Note that a PPT is defined only with respect to those α for

*Received by the editors December 12, 1996; accepted for publication (in revised form) by R. Cottle September 11, 1997; published electronically June 9, 1998.
http://www.siam.org/journals/simax/19-4/31381.html
†Indian Statistical Institute, 110, Nelson Manickam Road, Aminjikarai, Madras 600 029, India (gsrm@isimad.ernet.in).
‡Indian Statistical Institute, 7, SJS Sansanwal Marg, New Delhi 110 016, India (tps@isid.ernet.in).

which $det\, A_{\alpha\alpha} \neq 0$. By convention, when $\alpha = \emptyset$, $det\, A_{\alpha\alpha} = 1$ and $M = A$ (see [2]). Whenever we refer to PPTs, we mean the ones which are well defined.

We shall recall the definitions of some matrix classes that are relevant to this paper. Let $A \in \boldsymbol{R}^{n\times n}$. Then A is said to be a $\boldsymbol{P}$-matrix ($\boldsymbol{P_0}$-matrix) if all its principal minors are positive (nonnegative); if all principal minors of A are nonzero, then A is called a nondegenerate matrix; A is semimonotone ($\boldsymbol{E_0}$) if (q, A) has a unique solution for every $q > 0$; A is fully semimonotone ($\boldsymbol{E_0^f}$) if every PPT of A is in $\boldsymbol{E_0}$; A is copositive ($\boldsymbol{C_0}$) if $x^t Ax \geq 0$ for every $x \geq 0$; A is fully copositive ($\boldsymbol{C_0^f}$) if every PPT of A is in $\boldsymbol{C_0}$. For the definition of INS and Lipschitzian matrices see section 3.

In this article, we present some new results pertaining to three matrix classes, namely, (i) the class of adequate matrices introduced by Ingleton [4], (ii) the class of fully copositive matrices introduced by Murthy and Parthasarathy [7], and (iii) the class of INS-matrices introduced by Stone [13].

In the case of adequate matrices (see section 2), our main result is that a column adequate matrix is in $\boldsymbol{Q}$ (in $\boldsymbol{Q_0}$) if and only if it is completely-$\boldsymbol{Q}$ (completely-$\boldsymbol{Q_0}$). Characterization of completely-$\boldsymbol{Q_0}$ matrices in general is a complex problem [1]. Murthy and Parthasarathy [6, 7, 8] have shown that nonnegative matrices, symmetric copositive matrices, $\boldsymbol{C_0^f}$-matrices and Lipschitzian matrices are in $\boldsymbol{Q_0}$ if and only if they are completely-$\boldsymbol{Q_0}$.

Within the class of $\boldsymbol{C_0^f}$-matrices, we provide a sufficient condition under which a matrix will be in $\boldsymbol{P_0}$. As a corollary to this result, we provide an alternative proof of a result due to Murthy and Parthasarathy which states that $\boldsymbol{C_0^f} \cap \boldsymbol{Q_0}$-matrices are in $\boldsymbol{P_0}$. As another consequence of this result, we deduce that a bisymmetric $\boldsymbol{E_0^f}$-matrix A is positive semidefinite if, and only if, the rows and columns of $A + A^t$ corresponding to the zero diagonal entries are zero.

Last, we consider the class of INS-matrices and show that a nondegenerate INS-matrix must necessarily satisfy the *block* property. There are no constructive characterizations of Lipschitzian or INS-matrices. In [8], the authors showed that Lipschitzian matrices must necessarily satisfy the block property, and Stone [14] showed that Lipschitzian matrices are nondegenerate INS-matrices. We conjecture that block property is a characterization of Lipschitzian matrices. It is proven that the conjecture is true in the cases of nonnegative or nonpositive matrices and 2×2 matrices.

The results on adequate and $\boldsymbol{C_0^f}$-matrices are presented in section 2, and the results on INS- and Lipschitzian matrices are presented in section 3.

2. Results on adequate and $\boldsymbol{C_0^f}$-matrices. A number of matrix classes are invariant under principal pivoting; i.e., if a matrix is in class $\mathcal{C}$, then all its PPTs are also in $\mathcal{C}$. The matrix classes $\boldsymbol{Q}$, $\boldsymbol{Q_0}$, $\boldsymbol{P}$, $\boldsymbol{P_0}$, $\boldsymbol{E_0^f}$, $\boldsymbol{C_0^f}$, INS- and Lipschitzian matrices all fall in this category. In the definition below we consider another class of matrices which is also invariant under PPTs.

DEFINITION 2.1. *Say that a real square matrix $A \in \Lambda$ if for every PPT M of A the diagonal entries are nonnegative.*

Remark 2.2. Note that $\boldsymbol{E_0^f}$, which contains the classes $\boldsymbol{P_0}$ and $\boldsymbol{C_0^f}$ (see [2, 6, 7]), is a subclass of Λ. However, $\Lambda \setminus \boldsymbol{E_0^f}$ is nonempty as $\left[\begin{smallmatrix} 0 & -1 \\ -1 & 0 \end{smallmatrix}\right]$ is an example of this kind. Furthermore, it is easy to check that if $A \in \Lambda$, then $A^t \in \Lambda$.

Another class of matrices that is required for our results is the following.

DEFINITION 2.3. *Say that a real square matrix A has property (D) if for every index set α the following holds:*

$$\det A_{\alpha\alpha} = 0 \;\Rightarrow\; \text{columns of } A_{.\alpha} \text{ are linearly dependent.}$$

Let $\boldsymbol{\mathcal{D}}$ denote the class of matrices satisfying property (D). Note that if $A \in \Lambda$ ($A \in \boldsymbol{\mathcal{D}}$), then $A_{\alpha\alpha} \in \Lambda$ ($A_{\alpha\alpha} \in \boldsymbol{\mathcal{D}}$) for every α. An interesting property of $\boldsymbol{\mathcal{D}}$ is that if $A \in \boldsymbol{\mathcal{D}}$, then (q, A) has a solution with a complementary basis for any q with $S(q, A) \neq \phi$ (see [7]). Another interesting property of $\boldsymbol{\mathcal{D}}$, which is a direct consequence of the definition, is the following.

PROPOSITION 2.4. *If $A \in \boldsymbol{\mathcal{D}}$ is nonsingular, then A is nondegenerate.*

A matrix A is said to be a column (row) adequate matrix if A (A^t) is in $\boldsymbol{\mathcal{D}} \cap \boldsymbol{P_0}$. Ingleton [4] introduced the class of adequate matrices (i.e., both row and column adequate) and showed that if A is adequate, then, for every q with $S(q, A) \neq \phi$, $Az + q$ is unique over $S(q, A)$. We now present our main results on column adequate matrices.

THEOREM 2.5. *If $A \in \Lambda \cap \boldsymbol{\mathcal{D}}$, then $A \in \boldsymbol{P_0}$.*

Proof. We prove this by induction on n. Obviously the theorem is true if $n = 1$. Assume that the theorem is true for all $(n-1) \times (n-1)$ matrices. Let $A \in \boldsymbol{R}^{n\times n} \cap \Lambda \cap \boldsymbol{\mathcal{D}}$. By above observations, $A_{\alpha\alpha} \in \boldsymbol{P_0}$ for all α such that $|\alpha| = n - 1$. Suppose $A \notin \boldsymbol{P_0}$. Then $\det A < 0$. Note that A is almost $\boldsymbol{P_0}$. Since $A \in \Lambda$, diagonal entries of A^{-1} are equal to zero. This means that $\det A_{\alpha\alpha} = 0$ for all α with $|\alpha| = n - 1$. Since $A \in \boldsymbol{\mathcal{D}}$, this implies that columns of A are linearly dependent which contradicts that A is nonsingular. It follows that $A \in \boldsymbol{P_0}$. ∎

COROLLARY 2.6. *Suppose $A \in \boldsymbol{R}^{n\times n}$. The following conditions are equivalent:*

(a) $A \in \boldsymbol{P_0} \cap \boldsymbol{\mathcal{D}}$;

(b) $A \in \Lambda \cap \boldsymbol{\mathcal{D}}$.

It is known that nondegenerate $\boldsymbol{E_0^f}$-matrices are $\boldsymbol{P}$-matrices.

COROLLARY 2.7. *If $\boldsymbol{E_0^f} \cap \boldsymbol{\mathcal{D}}$, then $A \in \boldsymbol{P_0}$.*

A matrix A is said to be completely-$\boldsymbol{Q}$ (completely-$\boldsymbol{Q_0}$) if all its principal submatrices including A are $\boldsymbol{Q}$-matrices ($\boldsymbol{Q_0}$-matrices). Cottle introduced these classes in [1] and characterized completely-$\boldsymbol{Q}$ matrices as the class of strictly semimonotone matrices (A is said to be strictly semimonotone if (q, A) has a unique solution for every nonnegative q). One of the problems posed by Cottle [1] is the characterization of completely-$\boldsymbol{Q_0}$ matrices which is still an open problem. Murthy and Parthasarathy have characterized completely-$\boldsymbol{Q_0}$ matrices in certain special cases (see [6, 7, 8]). The following result augments these special cases with column adequate matrices.

THEOREM 2.8. *Suppose $A \in \Lambda \cap \boldsymbol{\mathcal{D}}$. Then*

(a) *$A \in \boldsymbol{Q_0}$ if and only if A is completely-$\boldsymbol{Q_0}$;*

(b) *$A \in \boldsymbol{Q}$ if and only if A is completely-$\boldsymbol{Q}$.*

Proof. (a) It suffices to show the "only if" part. Suppose $A_{\alpha\alpha} \notin \boldsymbol{Q_0}$, say, for $\alpha = \{1, 2, \ldots, n-1\}$. By Theorem 2.19 of [7], there exists a β such that $n \in \beta$, $\det A_{\beta\beta} \neq 0$ and $M_{.n} \leq 0$, where $M = \wp_\beta(A)$. Since $A \in \boldsymbol{P_0}$ (Theorem 2.5 above), $M_{nn} = \frac{\det A_{\gamma\gamma}}{\det A_{\beta\beta}} = 0$, where $\gamma = \beta \setminus \{n\}$. This implies $\det A_{\gamma\gamma} = 0$, which in turn implies $\det A_{\beta\beta} = 0$ as $A \in \boldsymbol{\mathcal{D}}$. From this contradiction, it follows that $A_{\alpha\alpha} \in \boldsymbol{Q_0}$. By induction it follows that A is completely-$\boldsymbol{Q_0}$.

(b) Once again, we will show the "only if" part. Note that the conclusions of Theorem 2.19 of [7] remain valid even if we replace $\boldsymbol{Q_0}$ by $\boldsymbol{Q}$ in the statement of that theorem (almost the same proof can be repeated). Hence it follows (from the proof of part (a) here) that A is completely-$\boldsymbol{Q}$. ∎

COROLLARY 2.9. *Every column adequate matrix is in* $\boldsymbol{Q}$ *if and only if it is strictly semimonotone.*

We now turn our attention to the results on $\boldsymbol{C_0^f}$-matrices. In [6], using the concept of *incidence*, it was shown that $\boldsymbol{C_0^f} \cap \boldsymbol{Q_0} \subseteq \boldsymbol{P_0}$. We recapture this result as a consequence of our results here.

THEOREM 2.10. *Suppose* $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C_0^f}$, $n \geq 2$. *If the rows and columns of* $A + A^t$ *corresponding to the zero diagonal entries of* A *are zero, then* $A \in \boldsymbol{P_0}$.

Proof. From the hypothesis and Theorem 3.17 of [7], it is clear that every 2×2 principal submatrix of A is in $\boldsymbol{P_0}$. Assuming that every $(k-1) \times (k-1)$, $k \geq 2$, principal submatrix of A is in $\boldsymbol{P_0}$, we will show that every $k \times k$ principal submatrix of A is also in $\boldsymbol{P_0}$. Let B be any $k \times k$ principal submatrix of A such that all its proper principal minors are nonnegative. Suppose $det\, B < 0$. Arguing as in Theorem 3.17 of [7], we can show that

$$B^{-1} = \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix},$$

where C and D are nonnegative square matrices of the same order. It follows that C and D are nonsingular and that $B = \left[\begin{smallmatrix} 0 & D^{-1} \\ C^{-1} & 0 \end{smallmatrix}\right]$. From the hypothesis, it follows that $C^{-1} + (D^{-1})^t = 0$ and hence $D^{-1} = -(C^{-1})^t$. This in turn implies that $D = -C^t$. This contradicts that D is nonnegative. Hence $det\, B \geq 0$. The theorem follows. ∎

COROLLARY 2.11. *Suppose* $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C_0^f} \cap \boldsymbol{Q_0}$. *Then* $A \in \boldsymbol{P_0}$.

Proof. If $n = 1$, there is nothing to prove. Assume $n \geq 2$. We will show that every 2×2 principal submatrix of A is in $\boldsymbol{P_0}$. Suppose, to the contrary, assume that $A_{\alpha\alpha} \notin \boldsymbol{P_0}$ for some α with $|\alpha| = 2$. Without loss of generality, we may take $\alpha = \{1, 2\}$. Then $A_{\alpha\alpha} \simeq \left[\begin{smallmatrix} 0 & + \\ + & 0 \end{smallmatrix}\right]$ (this notation means $a_{11} = a_{22} = 0$ and a_{12}, a_{21} are positive). Since $A_{\alpha\alpha} \notin \boldsymbol{Q_0}$, we must have $n > 2$ and a $j \in \bar{\alpha}$ such that $a_{j1} < 0$ (follows from Theorem 2.9 of [7]). Note that if $a_{1j} \leq 0$, then $A \notin \boldsymbol{C_0^f}$. But if $a_{1j} > 0$, then also $A \notin \boldsymbol{C_0^f}$ (follows from Theorem 4.1 of [8]). It follows that every 2×2 principal submatrix of A is in $\boldsymbol{P_0}$ and hence $A \in \boldsymbol{P_0}$. Arguing as in Lemma 3.2 of [6], we can show that for every i such that $a_{ii} = 0$, we have $a_{ij} + a_{ji} = 0$ for all j. Notice that in the proof of Lemma 3.2 of [6] we need only that every 2×2 principal submatrix of A is in $\boldsymbol{P_0}$. Hence the rows and columns of $A + A^t$ corresponding to zero diagonal entries of A are zero. From Theorem 2.10, it follows that $A \in \boldsymbol{P_0}$. ∎

In [6], it was shown that a $\boldsymbol{C_0^f}$-matrix is in $\boldsymbol{Q_0}$ if and only if it is completely-$\boldsymbol{Q_0}$. The arguments used to prove this can be extended to obtain the following result.

THEOREM 2.12. *Suppose* $A \in \boldsymbol{R}^{n\times n} \cap \boldsymbol{C_0^f}$. *If* $A \in \boldsymbol{Q_0}$, *then* A^t *and all its PPTs are completely-*$\boldsymbol{Q_0}$.

Proof. It can be verified that if a matrix $B \in \Lambda$ satisfies the condition that for every PPT C of B satisfies

$$c_{ii} = 0 \Rightarrow c_{ij} + c_{ji} = 0 \text{ for all } i \text{ and } j,$$

then B and all its PPTs are completely-$\boldsymbol{Q_0}$ matrices. This is because, if B has this property, then Graves's algorithm processes (q, B) for any q and terminates either with a solution or with the conclusion that $F(q, B) = \emptyset$ (see Chapter 4 of [10] and Theorem 3.4 of [6]). Therefore, we will show that any PPT of A^t will satisfy the above condition. Let $D = \wp_\alpha(A^t)$ for some α. Observe that $\wp_\alpha(A)$ exists. Let $M = \wp_\alpha(A)$. It can be checked that, $M = SD^tS$, where $S = \left[\begin{smallmatrix} I_{\alpha\alpha} & 0 \\ 0 & -I_{\bar{\alpha}\bar{\alpha}} \end{smallmatrix}\right]$. Hence for each i, j, either $d_{ij} + d_{ji} = m_{ij} + m_{ji}$ or $d_{ij} + d_{ji} = -(m_{ij} + m_{ji})$. If $d_{ii} = 0$ for some i, then $m_{ii} = 0$,

and by Theorem 3.4 of [6], $m_{ij} + m_{ji} = 0$. From this it follows that if for some i, $d_{ii} = 0$, then $d_{ij} + d_{ji} = 0$. □

One may ask whether the converse of the above theorem is true. That is, if $A \in \boldsymbol{C_0^f}$ and A^t and all its PPTs are completely-$\boldsymbol{Q_0}$, then is it true that $A \in \boldsymbol{Q_0}$? The answer to this question is "no." The problem arises from the fact that transpose of a $\boldsymbol{C_0^f}$-matrix need not be in $\boldsymbol{C_0^f}$. As a counter example, consider the $\boldsymbol{C_0^f}$-matrix $A = \left[\begin{smallmatrix} 1 & 0 \\ 1 & 0 \end{smallmatrix}\right]$. It can be checked, directly or using Theorem 2.5 of [7], that A^t and its PPT are completely-$\boldsymbol{Q_0}$ but $A \notin \boldsymbol{Q_0}$.

A matrix A is said to be bisymmetric if, for some index set α, $A_{\alpha\alpha}$ and $A_{\bar{\alpha}\bar{\alpha}}$ are symmetric and $A_{\bar{\alpha}\alpha} = -A^t_{\alpha\bar{\alpha}}$. It is easy to check that PPTs of bisymmetric matrices are bisymmetric.

THEOREM 2.13. *Suppose $A \in \boldsymbol{R}^{n\times n}$ is a bisymmetric $\boldsymbol{E_0^f}$-matrix. Then the following are equivalent:*

(a) $A \in \boldsymbol{Q_0}$;
(b) *A is positive semidefinite*;
(c) *for any i, j, $a_{ii} = 0 \;\Rightarrow\; a_{ij} + a_{ji} = 0$;*
(d) *every 2×2 principal submatrix of A is in $\boldsymbol{P_0}$.*

Proof. We first observe that every bisymmetric $\boldsymbol{E_0^f}$-matrix is in $\boldsymbol{C_0^f}$ (Theorem 4.7 of [6]). Implication (a) ⇒ (b) was already established in [6]. The implication (b) ⇒ (c) is a well-known fact about positive semidefinite matrices. The implication (c) ⇒ (d) is a direct consequence of Theorem 2.10. To complete the proof of the theorem, we will show that (d) ⇒ (a). Assume that A satisfies (d). Using the fact that every 2×2 principal submatrix of A is in $\boldsymbol{C_0^f} \cap \boldsymbol{P_0}$, it is easy to show that A satisfies (c). Hence, by 2.10, $A \in \boldsymbol{P_0}$. Let M be any PPT of A. Suppose $m_{ii} = 0$ for some i. As A is bisymmetric, so is M. So for any j, either $m_{ij} = -m_{ji}$ or $m_{ij} = m_{ji}$. If $m_{ij} = -m_{ji}$, then $m_{ij} + m_{ji} = 0$. If $m_{ij} = m_{ji}$, then, as $M \in \boldsymbol{P_0}$ and $m_{ii} = 0$, we must have $m_{ij} = m_{ji} = 0$. Thus for any j, $m_{ij} + m_{ji} = 0$. By Theorem 3.4 of [6], it follows that $A \in \boldsymbol{Q_0}$. □

3. Block property. Stone [13] introduced the class of INS-matrices. A matrix A is said to be an INS_k-matrix if $|S(q, A)| = k$ for all $q \in \text{int}K(A)$, where $K(A)$ is the set of all p for which $S(p, A) \neq \emptyset$; and $INS = \cup_{k=0}^{\infty} INS_k$. Next we say that A is Lipschitzian matrix if there exists a positive number λ, called the Lipschitzian constant, such that for any $p, q \in K(A)$, the following holds: given any $x \in S(p, A)$, there exists a $z \in S(q, A)$ such that $\|x - z\| \;\leq\; \lambda\|p - q\|$. Stone [14] showed that Lipschitzian matrices are nondegenerate INS-matrices and conjectured that the converse is also true. Furthermore, he showed that the conjecture is true with an additional assumption of Lipschitz *path-connectedness* (see [14] for details). To date, no constructive characterizations are known for INS and Lipschitzian matrix classes. Thus, there is no finite procedure to verify whether a given matrix is INS or Lipschitzian.

DEFINITION 3.1. *Say that A has property (B) if every PPT M of A has the following block structure (subject to a principal rearrangement):*

$$M = \begin{bmatrix} M_{11} & 0 & \dots & 0 & M_{1\overline{l+1}} \\ 0 & M_{22} & \dots & 0 & M_{2\overline{l+1}} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \dots & M_{ll} & M_{l\overline{l+1}} \\ M_{\overline{l+1}1} & M_{\overline{l+1}2} & \dots & M_{\overline{l+1}l} & M_{\overline{l+1}\,\overline{l+1}} \end{bmatrix},$$

where $M_{11}, M_{22}, \ldots, M_{ll}$ are all negative $\boldsymbol{N}$-matrices (i.e., all entries and all principal minors are negative) and the diagonal entries of $M_{\overline{l+1}\ \overline{l+1}}$ are positive.

In [8], the authors showed that every Lipschitzian matrix must have property (B). In this section, we will show that every nondegenerate INS-matrix also must have property (B).

Note that if a matrix A has property (B), then it must be nondegenerate as every PPT of A has no zero diagonal entries (see Corollary 3.5, p. 204 of [10]). From the definition, property (B) is invariant under PPTs and is inherited by all the principal submatrices.

THEOREM 3.2. *Suppose $A \in \boldsymbol{R}^{n\times n}$ is a nondegenerate INS-matrix. Then A has property (B).*

Proof. Let $\alpha = \{i : a_{ii} < 0\}$. By Theorem 5 of [12], $A_{\alpha\alpha}$ is a nondegenerate INS-matrix. Also, for $i, j \in \alpha,\ i \neq j,\ A_{\beta\beta} \in INS$, where $\beta = \{i, j\}$. It is easy to check that if $A_{\beta\beta}$ has a positive entry, then $A_{\beta\beta} \notin INS$. It follows that $A_{\alpha\alpha}$ is nonpositive and hence in $\boldsymbol{Q_0}$. From Corollary 3.5 of [14], $A_{\alpha\alpha}$ is Lipschitzian. From Theorem 4.7 of [8],

$$A_{\alpha\alpha} = \begin{bmatrix} N^1 & 0 & \ldots & 0 \\ 0 & N^2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & N^l \end{bmatrix} \text{ for some } l \geq 1,$$

where each N^i is a negative $\boldsymbol{N}$-matrix. Since every PPT of nondegenerate INS-matrix is also nondegenerate INS, we conclude that A has property (B). □

Our conjecture is that property (B) is also sufficient condition for a matrix to be Lipschitzian. Below we verify this conjecture in certain special cases.

THEOREM 3.3. *Suppose $A \in \boldsymbol{R}^{n\times n}$. Assume that any one of the following conditions holds:*

(i) $n = 2$;
(ii) $A \leq 0$;
(iii) *A is completely-$\boldsymbol{Q}$*;
(iv) $A \geq 0$.

Then the following statements are equivalent:

(a) *A is nondegenerate INS*;
(b) *A is Lipschitzian*;
(c) *A has property (B).*

Proof. In view of Stone's result that (b) ⇒ (a) (Theorem 3.2 of [14]), it suffices to show that (c) implies (b). So assume that (c) holds.

(i). If the diagonal entries of A are negative, then property (B) implies that either A is a negative $\boldsymbol{N}$-matrix or $A \simeq \left[\begin{smallmatrix} - & 0 \\ 0 & - \end{smallmatrix}\right]$. In either case, A is Lipschitzian (see [3]). If the diagonal entries of A are positive, then either A is a $\boldsymbol{P}$-matrix or A^{-1} is a negative $\boldsymbol{N}$-matrix. Once again A is Lipschitzian (see [5]). Consider the last case $a_{11} < 0$ and $a_{22} > 0$, without loss of generality. It is easy to check (graphically) that A is INS and that $K(A)$ is Lipschitz path-connected (see [14] for details and the example following Definition 3.3 in [14]). From Theorem 3.4 of [14], we conclude A is Lipschitzian.

(ii). By property (B), A can be decomposed into a block diagonal matrix where each submatrix on the diagonal is a negative $\boldsymbol{N}$-matrix. As negative $\boldsymbol{N}$-matrices are Lipschitzian, one can easily verify that A is also Lipschitzian.

(iii). In this case we actually show that A is a $\boldsymbol{P}$-matrix and this we do by induction on the order of the matrix. Obviously the result is true for $n = 1$. Assume

the result for all matrices of order $n-1,\ n>1$. Suppose $A \in \boldsymbol{R}^{n\times n}$ satisfies the hypothesis. Then all the proper principal minors of A are positive. If $A \notin \boldsymbol{P}$, then $det\, A < 0$ and the diagonal entries of A^{-1} are negative. By property (B), A^{-1} must be nonpositive. But this contradicts that $A \in \boldsymbol{Q}$. Hence $A \in \boldsymbol{P}$.

(iv). From the hypothesis and (c), $a_{ii} > 0$ for all i. Since $A \geq 0$, A is completely-$\boldsymbol{Q}$. Therefore $A \in \boldsymbol{P}$. □

PROPOSITION 3.4. *Suppose $A \in \boldsymbol{R}^{n\times n}$. Assume that for some index set α, $A_{\alpha\alpha}$ is Lipschitzian and $A_{\bar{\alpha}\bar{\alpha}} \in \boldsymbol{P}$. If $A_{\bar{\alpha}\alpha} = 0$ or $A_{\alpha\bar{\alpha}} = 0$, then A is Lipschitzian.*

Proof. Assume $A_{\bar{\alpha}\alpha} = 0$. Let $p, q \in K(A)$. Let λ_1 and λ_2 be the Lipschitzian constants corresponding to $A_{\alpha\alpha}$ and $A_{\bar{\alpha}\bar{\alpha}}$ respectively. Take any arbitrary $x \in S(p, A)$. We will exhibit a $z \in S(q, A)$ such that $\|z - x\| \leq \lambda\|p - q\|$, where λ, to be chosen later, depends only on λ_1, λ_2, and A. Since $S(q, A) \neq \phi$, choose any $\bar{z} \in S(q, A)$. Let $y = Ax + p$ and $\bar{w} = A\bar{z} + q$. Note that $x_\alpha \in S(p'_\alpha, A_{\alpha\alpha})$ and $\bar{z}_\alpha \in S(q'_\alpha, A_{\alpha\alpha})$, where $p'_\alpha = p_\alpha + A_{\alpha\bar{\alpha}}x_{\bar{\alpha}}$ and $q'_\alpha = q_\alpha + A_{\alpha\bar{\alpha}}\bar{z}_{\bar{\alpha}}$. Since $A_{\alpha\alpha}$ is Lipschitzian, there exists a $z_\alpha \in S(q'_\alpha, A_{\alpha\alpha})$ such that

$$\begin{aligned} \|x_\alpha - z_\alpha\| &\leq \lambda_1\|p'_\alpha - q'_\alpha\| \\ &\leq \lambda_1\|p_\alpha - q_\alpha\| + \lambda_1\|B\|\|x_{\bar{\alpha}} - z_{\bar{\alpha}}\|. \end{aligned}$$

Since $z_\alpha \in S(q'_\alpha, A_{\alpha\alpha})$, $w_\alpha = A_{\alpha\alpha}z_\alpha + q_\alpha + A_{\alpha\bar{\alpha}}\bar{z}_{\bar{\alpha}}$ and $w_\alpha^t z_\alpha = 0$. This implies $z = (z_\alpha^t, \bar{z}_{\bar{\alpha}}^t)^t \in S(q, A)$. As $A_{\bar{\alpha}\bar{\alpha}} \in \boldsymbol{P}$, $x_{\bar{\alpha}}$ and $z_{\bar{\alpha}}$ are the unique solutions of $(p_{\bar{\alpha}}, A_{\bar{\alpha}\bar{\alpha}})$ and $(q_{\bar{\alpha}}, A_{\bar{\alpha}\bar{\alpha}})$. Therefore, $\|x_{\bar{\alpha}} - z_{\bar{\alpha}}\| \leq \lambda_2\|p_{\bar{\alpha}} - q_{\bar{\alpha}}\|$. Combining this with the above inequality, we get

$$\begin{aligned} \|x - z\| &\leq \|x_\alpha - z_\alpha\| \\ &\leq \lambda_1\|p_\alpha - q_\alpha\| + (\lambda_1\lambda_2\|B\| + \lambda_2)\|p_{\bar{\alpha}} - q_{\bar{\alpha}}\| \\ &\leq \lambda_1\|p - q\| + (\lambda_1\lambda_2\|B\| + \lambda_2)\|p - q\| \\ &\leq \lambda\|p - q\|, \text{ where } \lambda = \lambda_1 + \lambda_2 + \lambda_1\lambda_2\|B\|. \end{aligned}$$

It follows that A is Lipschitzian, and the case $A_{\alpha\bar{\alpha}} = 0$ can be tackled in a similar fashion. □

Proposition 3.4 is not valid if we simply assume that $A_{\alpha\alpha}$ and $A_{\bar{\alpha}\bar{\alpha}}$ are Lipschitzian. As a counter example, consider $A = \left[\begin{smallmatrix} -1 & 1 \\ 0 & -1 \end{smallmatrix}\right]$. It is clear that $A \notin INS$, and hence A is not Lipschitzian.

The following is an example of a matrix with property (B).

Example 3.5.

$$A = \begin{bmatrix} -1 & -2 & -2 \\ -2 & -1 & 1 \\ 1 & 0 & 1 \end{bmatrix}.$$

It is not known whether A is Lipschitzian or not.

Acknowledgments. The authors wish to thank Dr. G. Ravindran and Mr. Amit K. Biswas for some useful discussions.

REFERENCES

[1] R. W. COTTLE, *A note on completely $\mathcal{Q}$-matrices,* Math. Programming, 9 (1980), pp. 347–351.
[2] R. W. COTTLE, J. S. PANG, AND R. E. STONE, *The Linear Complementarity Problem,* Academic Press, New York, 1992.

[3] M. S. Gowda, *On the continuity of solution map in linear complementarity problems,* SIAM J. Optim., 2 (1992), pp. 619–634.

[4] A. W. Ingleton, *A problem in linear inequalities,* Proc. London Math. Soc., 16 (1966), pp. 519–536.

[5] O. L. Mangasarian and T. H. Shiau, *Lipschitz continuity of solutions of linear inequalities, programs and complementarity problems,* SIAM J. Control Optim., 25 (1987), pp. 583–595.

[6] G. S. R. Murthy and T. Parthasarathy, *Fully copositive matrices,* Math. Programming, to appear.

[7] G. S. R. Murthy and T. Parthasarathy, *Some properties of fully semimonotone $\boldsymbol{Q_0}$-matrices,* SIAM J. Matrix Anal. Appl., 16 (1995), pp. 1268–1286.

[8] G. S. R. Murthy, T. Parthasarathy, and B. Sriparna, *Constructive characterization of Lipschitzian $\boldsymbol{Q_0}$-matrices,* Linear Algebra Appl., 252 (1997), pp. 323–337.

[9] K. G. Murty, *On the number of solutions to the complementarity problem and spanning properties of complementary cones,* Linear Algebra Appl., 5 (1972), pp. 65–108.

[10] K. G. Murty, *Linear Complementarity, Linear and Nonlinear Programming,* Heldermann Verlag, Berlin, Germany, 1988.

[11] T. D. Parsons, *Applications of principal pivoting,* in Proceedings of the Princeton Symposium on Mathematical Programming, (H. W. Kuhn, ed., Princeton University Press, Princeton, NJ), 1970, pp. 561–581.

[12] T. Parthasarathy, R. Sridhar, and B. Sriparna, *On Lipschitzian $\boldsymbol{Q_0}$ and INS-matrices,* Linear Algebra Appl., 263 (1997), pp. 193–199.

[13] R. E. Stone, *Geometric aspects of linear complementarity problem,* Ph.D. thesis, Department of Operations Research, Stanford University, Stanford, CA, 1981.

[14] R. E. Stone, *Lipschitzian matrices are nondegenerate INS-matrices,* in Complementarity and Variational Problems: State of the Art, M.C. Ferris and J.S. Pang, eds., SIAM, Philadelphia, 1997, pp. 440–451.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 906–923, October 1998

006

STATISTICAL CONDITION ESTIMATION FOR LINEAR LEAST SQUARES*

C. S. KENNEY†, A. J. LAUB†, AND M. S. REESE†

Abstract. Statistical condition estimation is applied to the linear least squares problem. The method obtains componentwise condition estimates via the Fréchet derivative. A rigorous statistical theory exists that determines the probability of accuracy in the estimates. The method is as computationally efficient as normwise condition estimation methods, and it is easily adapted to respect structural constraints on perturbations of the input data. Several examples illustrate the method.

Key words. conditioning, sensitivity, linear least squares

AMS subject classifications. 15A06, 15A12, 62J05, 65F05, 65F20, 65F30, 65F35

PII. S0895479895291935

1. Introduction.

The linear least squares problem

$$\min_x \|Ax - b\|_2,$$

where $b \in \mathbb{R}^m$ and $A \in \mathbb{R}_q^{m\times n}$ (subscript on $\mathbb{R}$ is the rank of A), has unique minimum 2-norm solution

$$x = A^+ b, \tag{1}$$

where A^+ is the Moore–Penrose pseudoinverse of A. Standard approaches [8, Chapter 9], [11], [13, section 4], [27] to determining the sensitivity of the solution to the linear least squares problem estimate the 2-norm condition number with respect to pseudoinversion of A. This value, denoted $\kappa_2(A)$, is given by

$$\kappa_2(A) = \|A\|_2\|A^+\|_2 = \frac{\sigma_1(A)}{\sigma_q(A)},$$

where $\sigma_1(A) \geq \cdots \geq \sigma_q(A) > 0$ are the nonzero singular values of A. Indeed, $\kappa_2(A)$ plays a major role in bounding relative changes in the solution to a linear least squares problem due to perturbations in the arguments. For example, if $A, E \in \mathbb{R}^{m\times n}$ such that A has full rank and $\|A^+\|_2\|E\|_2 < 0.2$, and if $x \neq 0$ solves

$$\min_z \|Az - b\|_2$$

and $x + y$ solves

$$\min_z \|(A + E)z - b\|_2,$$

then [8, p. 9.5] $A + E$ has full rank and

$$\frac{\|y\|_2}{\|x\|_2} \leq 1.6\left(\kappa_2(A) + \kappa_2^2(A)\frac{\|b - Ax\|_2}{\|A\|_2\|x\|_2}\right)\frac{\|E\|_2}{\|A\|_2}. \tag{2}$$

* Received by the editors September 18, 1995; accepted for publication (in revised form) by P. Van Dooren July 12, 1997; published electronically June 9, 1998. This research was supported in part by National Science Foundation grant ECS-9633326, Air Force Office of Scientific Research grant F49620-94-1-0104DEF, and Office of Naval Research grant N00014-96-1-0456.

http://www.siam.org/journals/simax/19-4/29193.html

† Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106-9560 (laub@ucdavis.edu).

Note that the sensitivity of the linear least squares solution is proportional to $\kappa_2(A)$ when the residual $b - Ax$ is small compared with the norms of A and x; otherwise, it is proportional to the square of $\kappa_2(A)$ (also see [10, section 5.3.9]).

Normwise condition estimation approaches consolidate all sensitivity information into a single number. Thus, important information may be lost if individual solution components have widely disparate sensitivity. A method that estimates the componentwise condition of the solution vector is often preferable [1, section 4.3.2], [4], [5], [13], [21]. This is illustrated by the following example.

Example 1. Let

$$A = \begin{bmatrix} 1 & 1 \\ \epsilon & 0 \\ 0 & \epsilon \end{bmatrix}, \qquad x = \begin{bmatrix} 1/\epsilon \\ \epsilon \end{bmatrix}, \qquad \text{and} \qquad b = \begin{bmatrix} 1/\epsilon + \epsilon \\ 1 \\ \epsilon^2 \end{bmatrix},$$

where $\epsilon > 0$. In exact arithmetic, x is the minimum 2-norm solution to the linear least squares problem $\min_z \|Az - b\|_2$. However, suppose we are solving for x on a finite-word-length computer. Say we have 16 digits of precision, and let $\epsilon = 10^{-8}$. Then

$$A = \begin{bmatrix} 1 & 1 \\ 10^{-8} & 0 \\ 0 & 10^{-8} \end{bmatrix}, \qquad \overline{x} = \begin{bmatrix} 10^8 \\ 5 \times 10^{-9} \end{bmatrix}, \qquad \text{and} \qquad \overline{b} = \begin{bmatrix} 10^8 \\ 1 \\ 10^{-16} \end{bmatrix},$$

where $\overline{b}$ is b rounded to 16 digits and $\overline{x}$ is the exact minimum 2-norm solution to the linear least squares problem $\min_z \|Az - \overline{b}\|_2$ rounded to 16 digits. The 2-norm condition number of A is $\kappa_2(A) \approx 10^8$. The normwise relative error

$$\frac{\|x - \overline{x}\|_2}{\|x\|_2} \approx 5 \times 10^{-17}$$

is quite small; however, the componentwise relative errors are

$$\frac{|x_1 - \overline{x}_1|}{|x_1|} = 0,$$
$$\frac{|x_2 - \overline{x}_2|}{|x_2|} = 0.5.$$

The first component of the computed solution exhibits 0% relative error, while the second component exhibits 50% relative error. All the digits of $\overline{x}_1$ are correct, while the first digit of $\overline{x}_2$ is only correct upon rounding. Thus, individual components of the solution vector may be much more relatively ill conditioned than is indicated by the standard 2-norm condition number.

Things may also be better than they seem, i.e., $\kappa_2(A)$ may be a gross overestimate of the sensitivity of the problem, especially if b is well aligned with the range of A, denoted $\mathcal{R}(A)$. Examples in this paper were computed with MATLAB 4.2a on a Sun SPARCstation, which has a relative machine precision of `eps` $\approx 2.22 \times 10^{-16}$.

Example 2. Let A be as in Example 1, and let

$$b = \begin{bmatrix} 2 \\ \epsilon \\ \epsilon \end{bmatrix} \qquad \text{and} \qquad c = \begin{bmatrix} 1 + \epsilon \\ -1/\epsilon \\ -1/\epsilon \end{bmatrix}.$$

Note that $b \in \mathcal{R}(A)$ for all ϵ, while for small ϵ, say 10^{-8}, the vector c is almost orthogonal to $\mathcal{R}(A)$. MATLAB gives minimum 2-norm solutions $\overline{x}_b$ and $\overline{x}_c$ for the linear least squares problems corresponding to (A, b) and (A, c), respectively. If we denote the exact minimum 2-norm solutions by x_b and x_c, we find that

$$\frac{\|x_b - \overline{x}_b\|_2}{\|x_b\|_2} \approx 4 \times 10^{-16},$$
$$\frac{\|x_c - \overline{x}_c\|_2}{\|x_c\|_2} \approx 3 \times 10^{8}.$$

In fact, the relative error in each component of $\overline{x}_b$ is on the order of 10^{-16}, while the relative errors in the components of $\overline{x}_c$ are on the order of 1 and 10^8, respectively. Given the large condition number $\kappa(A) \approx 10^8$, we might be surprised that $\overline{x}_b$ is so close to x_b.

Finally, standard condition estimation methods do not necessarily respect the structure of certain problems.

Example 3. Suppose

$$A = \begin{bmatrix} \epsilon & 1 \\ 2\epsilon & 0 \\ 3\epsilon & 0 \end{bmatrix} \qquad \text{and} \qquad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Note that for $\epsilon = 10^{-8}$, we have $\kappa_2(A) \approx 3 \times 10^7$. However, the solution vector $x = A^+ b$ is sensitive only to perturbations of the first column of A; that is, if we restrict perturbations of A and b to be additive of the form

$$E = \begin{bmatrix} 0 & \epsilon_1 \\ 0 & \epsilon_2 \\ 0 & \epsilon_3 \end{bmatrix} \qquad \text{and} \qquad f = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \end{bmatrix},$$

respectively, where the ϵ_i and ϕ_i are small, then the linear least squares problem is quite well conditioned.

These examples illustrate three potential deficiencies in using $\kappa_2(A)$ to estimate the sensitivity of the solution to the linear least squares problem $\min_x \|Ax - b\|_2$.

1. The sensitivity of each component of the solution vector is not calculated; hence, $\kappa_2(A)$ may grossly underestimate the sensitivity of specific components of the solution vector.
2. If perturbations are due to rounding and b is well aligned with $\mathcal{R}(A)$, then $\kappa_2(A)$ may grossly overestimate the sensitivity of the solution.
3. When perturbations are restricted to a specific structure, $\kappa_2(A)$ may grossly overestimate the sensitivity of the solution.

Introduced by Kenney and Laub in [16], a statistically based method for estimating the condition of general matrix functions addresses these deficiencies. The method produces componentwise condition estimates for matrix functions in such a way that all input data can be considered in forming the estimates. This is done by evaluating a matrix function at the original arguments as well as at slightly perturbed arguments. Condition is then estimated by considering the effect on the solution due to perturbations of the arguments. A small number of function evaluations at perturbed arguments suffices to give a highly reliable condition estimate; hence, the method is referred to as small-sample statistical condition estimation (SCE). For general matrix functions, the amount of computation is directly proportional to the number of

function evaluations done, similar to Stewart's stochastic perturbation theory [24]. However, for many problems (in particular, linear systems and linear least squares), SCE is able to exploit Fréchet derivatives and factorizations of the data to reduce the computational cost to that of standard condition estimation methods. SCE has accuracy and storage requirements similar to other condition estimation approaches, and it is easy to adapt the method to respect constraints on the structure of allowable perturbations.

In his comprehensive survey of componentwise perturbation results [13], Higham defines the componentwise relative backward error of the computed solution $\overline{x}$ to a linear system $Ax = b$ as

$$\omega_{|A|,|b|}(\overline{x}) = \min\{\epsilon : (A+E)\overline{x} = b + f,\ |E| \le \epsilon|A|,\ |f| \le \epsilon|b|\},$$

with matrix absolute values being defined componentwise and inequalities holding componentwise. This scalar is the same as the output value `BERR` from the LAPACK [1] expert driver routines `xyySVX`. It is essentially a bound on the relative perturbations in components of A and b such that $\overline{x}$ is an exact solution. SCE, on the other hand, provides a full vector of condition numbers, giving an estimate of the sensitivity of each entry in the solution, similar to the approach of Chandrasekaran [4] and Chandrasekaran and Ipsen [5]. Furthermore, LAPACK assumes that perturbations have magnitudes relative to the input data, while SCE and Higham's analysis allow more general perturbations.

The application of SCE to nonsingular linear systems was investigated in [17]. In the present paper we examine how SCE can be applied to (full column rank) linear least squares problems. The remainder of this section introduces notation and reviews the theories of SCE and linear least squares. Section 2 discusses how to apply SCE methods to unstructured linear least squares problems. There we re-examine Example 1 using SCE methods. Sections 3 and 4 apply SCE to linear least squares problems in which the perturbations are relative and structured, respectively. These sections contain examples and methods that illustrate various points in the discussion, including Examples 2 and 3. Section 5 compares computational costs of the SCE method and standard condition estimation methods. Section 6 illustrates the application of SCE to some well-known linear least squares examples.

1.1. Notation. For simplicity, our matrices and vectors have real entries; however, the theory can be extended easily to matrices and vectors with complex entries. Single vertical bars around a matrix or vector indicate the componentwise absolute value of the matrix or vector. A matrix or vector surrounded by single vertical bars and then raised to a power signifies that the absolute values of the entries are to be raised to that power. For example, if A is the diagonal matrix with diagonal entries $(-1, 2, -3)$, then $|A|^2$ is the diagonal matrix with diagonal entries $(1, 4, 9)$.

The vec operation forms the Kronecker vector of a matrix by stacking its columns. The unvec operation undoes the vec operation. Hence, if A is the 4×4 identity matrix I_4, then $v = \mathrm{vec}(A) = [e_1^T, e_2^T, e_3^T, e_4^T]^T$ and, assuming it is known from the context that $A \in \mathbb{R}^{4\times 4}$, then $A = \mathrm{unvec}(v)$ sets A equal to I_4.

If a vector z is selected uniformly and randomly from the unit sphere S_{p-1} in $\mathbb{R}^p$, we write $z \sim \mathrm{U}(S_{p-1})$. If scalars x_i are independently selected from a normal distribution with mean μ and variance σ^2, we write $x_i \sim \mathrm{N}(\mu, \sigma^2)$; for example, samples taken from the standard normal distribution are $\mathrm{N}(0, 1)$.

We denote the Fréchet derivative of a matrix function $g : \mathbb{R}^n \to \mathbb{R}^m$ with respect to the variable $X \in \mathbb{R}^p$ by $Dg(X) = (\partial g_i/\partial x_j) \in \mathbb{R}^{m\times p}$. The Fréchet derivative of g

with respect to X evaluated in the direction $Y \in \mathbb{R}^p$ is denoted $Dg(X;Y) = Dg(X)Y$. When X and Y are matrices, we replace them by their Kronecker vectors in the derivative.

1.2. Review of statistical condition estimation. A function is said to be locally sensitive at a point if small changes in its arguments at that point can cause large relative changes in its value. The work in [16] shows that the local sensitivity (or condition) of a function at a point can be estimated accurately by randomly perturbing the arguments to the function and observing the effects on the function's value. In fact, [16, section 2] gives a rigorous theory for evaluating the probability of accuracy in the sensitivity estimate.

For example, suppose $g : \mathbb{R}^p \to \mathbb{R}$ is at least twice continuously differentiable. If we denote the gradient of g by the row vector v^T, then the local sensitivity of g at a point $x \in \mathbb{R}^p$ can be measured by the 2-norm of v evaluated at x. Expanding g in a Taylor series about x,

$$g(x + \delta z) = g(x) + \delta v(x)^T z + \mathrm{O}(\delta^2), \tag{3}$$

where $\delta \in \mathbb{R}$ is small and $z \in \mathbb{R}^p$ has unit 2-norm, we see that if $\|v\|_2$ is large, a small perturbation in x can yield a large change in g in the direction of z. Indeed, we see from (3) that $\|v\|_2$ is a first-order bound on the ratio of the function value error to the argument error

$$\frac{|g(x + \delta z) - g(x)|}{\|(x + \delta z) - x\|_2} \leq \|v\|_2 + \mathrm{O}(\delta).$$

The discussion in [16, section 2] shows that if $z \sim \mathrm{U}(S_{p-1})$, then the absolute value of the Newton quotient

$$dz \equiv \frac{g(x + \delta z) - g(x)}{\delta}$$

divided by the Wallis factor

$$\omega_p = \begin{cases} 1 & \text{for } p = 1, \\ \frac{2}{\pi} & \text{for } p = 2, \\ \frac{1 \cdot 3 \cdot 5 \cdots (p-2)}{2 \cdot 4 \cdot 6 \cdots (p-1)} & \text{for odd } p > 2, \\ \frac{2}{\pi} \frac{2 \cdot 4 \cdot 6 \cdots (p-2)}{3 \cdot 5 \cdot 7 \cdots (p-1)} & \text{for even } p > 2 \end{cases} \tag{4}$$

is a first-order condition estimator; that is, the probability of a relative error in the estimate is inversely proportional to the size of the error. This is because the expected value of the condition estimator

$$\nu \equiv \frac{|v^T z|}{\omega_p}$$

is equal to the 2-norm of v

$$E(\nu) = \|v\|_2,$$

and for $\gamma > 1$

$$\Pr(\|v\|_2/\gamma \leq \nu \leq \gamma \|v\|_2) \geq 1 - \frac{2}{\pi\gamma} + \mathrm{O}\left(\frac{1}{\gamma^2}\right).$$

In practice, the Wallis factor can be approximated accurately [16, section 6] by

$$\omega_p \approx \sqrt{\frac{2}{\pi(p-\frac{1}{2})}}. \tag{5}$$

For linear least squares problems, the subject of this paper, it is important to realize that we can avoid the approximation inherent in using the Newton quotient by evaluating the Fréchet derivative directly and efficiently from the QR factors of the data matrix (see (16)–(17)).

We can improve our estimator with more function evaluations. Suppose $\nu_1, \nu_2, \ldots, \nu_k \in \mathbb{R}$ are condition estimates corresponding to orthonormal vectors $z_1, z_2, \ldots, z_k \sim \mathrm{U}(S_{p-1})$. The original theoretical work on SCE by Kenney and Laub [16] discusses how such a set of vectors might be obtained, for example, by a QR decomposition of an arbitrary set of vectors $w_1, w_2, \ldots, w_k$ with $w_i \sim \mathrm{U}(S_{p-1})$. The expected value of the norm of the projection of v onto the span of the vectors z_i is

$$\mathrm{E}\left(\sqrt{|v^T z_1|^2 + \cdots + |v^T z_k|^2}\right) = \mathrm{E}\left(\sqrt{(\omega_p \nu_1)^2 + \cdots + (\omega_p \nu_k)^2}\right) = \frac{\omega_p}{\omega_k}\|v\|_2, \tag{6}$$

where ω_p and ω_k are defined in (4). We see from (6) that the subspace condition estimator

$$\nu(k) \equiv \frac{\omega_k}{\omega_p}\sqrt{|v^T z_1|^2 + \cdots + |v^T z_k|^2} \tag{7}$$

has expected value $\|v\|_2$. Thus, it is a kth-order condition estimator (see [16, section 2]); that is, the probability of a relative error of size γ in the condition estimate is proportional to γ^{-k}. For example, for $k = 3$, the estimator $\nu(3)$ has probability 0.9989 of being within an order of magnitude of the true condition number $\|v\|_2$. Relative accuracy within an order of magnitude is usually sufficient for estimating the local condition of a function. The averaged condition estimator

$$\zeta(k) = \frac{|v^T z_1| + \cdots + |v^T z_k|}{k\,\omega_p}, \tag{8}$$

which is the standard Monte Carlo method for finding the expected value (see [16, section 1]), is also a kth-order condition estimator, and it can be computed somewhat faster than the subspace estimator $\nu(k)$ (see section 5). For the averaged condition estimator, the vectors $z_i \sim \mathrm{U}(S_{p-1})$ are selected independently but are not necessarily mutually orthogonal.

To this point in the paper, the function g has been scalar valued; however, we can easily extend SCE to vector- and matrix-valued functions by viewing g as a map from $\mathbb{R}^p$ to $\mathbb{R}^q$ (using the operations vec and unvec to convert between matrices and vectors), where each of the q entries of g is a scalar-valued function. Evaluating the matrix function at a slightly perturbed argument yields a local condition estimate for each component of the computed solution.

1.3. Review of linear least squares. The local approximation of a nonlinear real-world problem by a linear model can give rise to discrepancies between predicted and actual data. Measurement error and computational rounding error can also contribute to the inexactness of the model. Thus, the linear model may be inconsistent. It also may be overdetermined or underdetermined. Even in the case of a consistent

linear model with a square system matrix, the matrix may be singular. Standard nonsingular linear system condition estimation methods [6], [12], [14] cannot handle these general situations; rather, we require linear least squares methods [3], [8], [10], [18], [25].

1.3.1. Perturbed linear least squares. Suppose $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. We assume for the remainder of this paper that neither A nor b is zero. The linear least squares problem is to find an $x \in \mathbb{R}^n$ such that Ax is as close as possible to b in the 2-norm sense, i.e., x is a solution to $\min_z \|Az - b\|_2$. In the unperturbed case, the unique solution is given by (1). However, due to measurement inaccuracies and finite-precision representation, A and b may be slightly perturbed from their true values [22], [23], in which case we say that A and b are subject to input error (see [1, section 4.1]). Rounding error may also occur during computation. We can treat both input errors and rounding errors as perturbations of the true input data.

Perturbations may be additive or multiplicative. Multiplicative perturbations are scaled componentwise by the original data entries, hence we also refer to them as relative perturbations (also see [1, section 4.3.2]). Multiplicative perturbations are the most common perturbations seen in practice, resulting naturally from both input errors (except for zero entries) and rounding errors. Suppose $A \in \mathbb{R}_n^{m\times n}$ and $b \in \mathbb{R}^m$, and define

$$\tilde{A} = A + E \qquad \text{and} \qquad \tilde{b} = b + f, \tag{9}$$

where E and f are perturbations. The perturbed data components take the form

$$\tilde{a}_{ij} = a_{ij} + e_{ij} \qquad \text{and} \qquad \tilde{b}_i = b_i + f_i.$$

In the case of relative perturbations, the entries of E and f are of the form

$$e_{ij} = a_{ij}\epsilon_{ij} \qquad \text{and} \qquad f_i = b_i\phi_i, \tag{10}$$

respectively, leading to perturbed data components of the form

$$\tilde{a}_{ij} = a_{ij}(1 + \epsilon_{ij}) \qquad \text{and} \qquad \tilde{b}_i = b_i(1 + \phi_i). \tag{11}$$

The goal of linear least squares in the perturbed case is to find $\tilde{x} = x + y$ to minimize

$$\|\tilde{A}\tilde{x} - \tilde{b}\|_2. \tag{12}$$

The unique minimum 2-norm solution to the perturbed linear least squares problem is $\tilde{x} = \tilde{A}^+\tilde{b}$. In (12) we have allowed perturbations in both A and b. This is reminiscent of the total least squares problem in which there is no underlying assumption that errors occur only in b. The total least squares problem is discussed extensively in [9] and [26].

The rank of a matrix may change when the matrix is perturbed. We say that the matrix E is an *acute perturbation* of the matrix A if $\operatorname{rank}(A+E) = \operatorname{rank}(A)$. The set of acute perturbations of A is the set on which the pseudoinverse is continuous about A. Suppose that E is an acute perturbation of A. Let

$$\mu = \max\left\{ \frac{\|E\|_2}{\|A\|_2}, \frac{\|f\|_2}{\|b\|_2} \right\},$$

and assume that $\mu\,\kappa_2(A) < 1$. Thus, μ is a measure of the maximum relative perturbation in the arguments to the linear least squares problem. Define

$$\theta = \sin^{-1}\left(\frac{\|b - Ax\|_2}{\|b\|_2}\right),$$

where x solves the least squares problem exactly. The angle θ measures the amount of misalignment between b and $\mathcal{R}(A)$. Note that if b is not orthogonal to $\mathcal{R}(A)$, then $x \neq 0$. In this case, [10, section 5.3.8] shows that

$$\frac{\|y\|_2}{\|x\|_2} \leq \frac{\mu}{\cos(\theta)}\left(2\kappa_2(A) + \sin(\theta)\kappa_2(A)^2\right) + \mathrm{O}(\mu^2),$$

where $y = \tilde{x} - x$ is the difference between the perturbed and exact solutions.

Thus, if b is well aligned with $\mathcal{R}(A)$, then the bound on the condition of the linear least squares solution is proportional to $\kappa_2(A)$, while if b is not well aligned with $\mathcal{R}(A)$, the bound is proportional to the square of $\kappa_2(A)$ divided by the cosine of θ. Note that if b is nearly orthogonal to $\mathcal{R}(A)$, then the bound may be extremely high. This echoes the result in (2).

1.3.2. The Fréchet derivative and linear least squares. Suppose that $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Perturbing $[A, b]$ to $[A + \delta E, b + \delta f]$ in the normal equations $A^TAx = A^Tb$, where $\delta \in \mathbb{R}$ and $[E, f]$ has Frobenius norm equal to one, we get the normal equations corresponding to the perturbed linear least squares problem

$$(A + \delta E)^T(A + \delta E)(x + \delta y_\delta) = (A + \delta E)^T(b + \delta f). \tag{13}$$

Express the solution to the linear least squares problem as a function of $[A, b]$

$$x = g([A, b]) = A^+b, \tag{14}$$

and let $y = \mathrm{D}g([A, b]; [E, f])$ be the Fréchet derivative of (14) with respect to $[A, b]$ evaluated in the direction $[E, f]$. After some algebraic manipulation of (13) and taking the limit as $\delta \to 0$, we find that y is a solution to the nonnegative definite linear system (see also [10, section 5.3.8] and [13, Theorem 4.2])

$$A^TAy = A^T(f - Ex) + E^T(b - Ax).$$

For nonsingular linear systems in which $Ax = b$, this reduces to the Fréchet derivative presented in [17, section 2]. If A has full column rank and we have its QR factors

$$QA = \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix}, \quad A = \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{15}$$

then

$$y = R^{-1}\left(Q_1(f - Ex) + R^{-T}E^T(b - Ax)\right) \tag{16}$$

$$= R^{-1}\left(Q_1(f - Ex) + R^{-T}E^TQ_2^TQ_2b\right). \tag{17}$$

In practice, we cannot obtain the exact Fréchet derivative since we do not have the exact solution x. However, we usually only need condition estimates to be within an order of magnitude or so of the true condition of the problem, hence an approximate solution $\overline{x}$ will give a sufficiently good approximation of the Fréchet derivative.

Suppose now that E and f are random variables whose entries are i.i.d. $\mathrm{N}(0,1)$, but $[E,f]$ does not necessarily have Frobenius norm equal to one. We can rewrite (16) as

$$y = \frac{1}{\|[E,f]\|_F} R^{-1}\left(g - Sx + \|b - Ax\|_2 R^{-T} h\right), \tag{18}$$

where

$$\begin{aligned} S &= Q_1 E \in \mathbb{R}^{n\times n}, \\ g &= Q_1 f \in \mathbb{R}^n, \\ h &= \frac{1}{\|b-Ax\|_2} E^T(b-Ax) \in \mathbb{R}^n . \end{aligned} \tag{19}$$

Since the rows of Q_1 and Q_2 are orthonormal and the residual $b - Ax$ lies in $\mathcal{R}(Q_2)$, the entries of S, g, and h are i.i.d. $\mathrm{N}(0,1)$ (see [15, p. 135] or [20] for details). Hence, S, g, and h can be produced directly rather than via the costly products of (19). Since we no longer have E and f, we replace $\|[E,f]\|_F$ in (18) by its expected value [7]

$$\mathrm{E}\left(\|[E,f]\|_F\right) = \frac{\sqrt{2}\,\Gamma\left(\frac{m(n+1)+1}{2}\right)}{\Gamma\left(\frac{m(n+1)}{2}\right)} \approx \sqrt{m(n+1)}.$$

Fréchet derivative information in the general case of non-full-rank A is more problematic. The easily derived identity (e.g., see [25])

$$\begin{aligned} (A+F)^+ - A^+ \equiv &-(A+F)^+ F A^+ + (A+F)^+ (A+F)^{+T} F^T (I - AA^+) \\ &+ (I - (A+F)^+(A+F)) F^T A^{+T} A^+ \end{aligned}$$

can be applied to derive an expression for $(A+\delta E)^+(b+\delta f) - A^+ b$ with notation as in the beginning of this section. However, there are two principal difficulties. The first is the need to compute or estimate $(A+\delta E)^+$, which cannot be determined directly or efficiently from an already computed QR factorization of A. The second and more important reason is that the pseudoinverse is, in general, a discontinuous function of its argument. Specifically, it is easily shown that if $\mathrm{rank}(A+\delta E) > \mathrm{rank}(A)$, and this is the generic situation if A is rank deficient, then

$$\|(A+\delta E)^+ - A^+\|_2 \geq \frac{1}{\delta}.$$

Thus, we shall confine our attention in this paper to the important practical case where $A \in \mathbb{R}_n^{m\times n}$.

2. SCE for full column rank linear least squares. We now combine the results of section 1.2 and (18) to obtain an SCE-based method for estimating the condition of the solutions to full-column-rank linear least squares problems. Inputs to the method are the matrix $A \in \mathbb{R}_n^{m\times n}$ and the vector $b \in \mathbb{R}^m$, and the output is the relative condition vector $\kappa_{\mathrm{rel}} \in \mathbb{R}^n$, which is an estimate of the relative sensitivity of each entry of the computed solution vector x. The method requires the QR factors of A, where $Q \in \mathbb{R}^{m\times m}$ is orthogonal and $R \in \mathbb{R}^{n\times n}$ is upper triangular. Generally, these will have been calculated during the solution of the linear least squares problem. The perturbations S_j, g_j, and h_j in the method correspond to S, g, and h of (18).

The integer $k \geq 1$ refers to the number of perturbations of input data. Note that when $k = 1$, there is no need to orthonormalize the set of vectors in Step 1 of the method.

ALGORITHM 1 (subspace condition estimation for linear least squares).

1. *Generate* $(S_1, g_1, h_1), (S_2, g_2, h_2), \ldots, (S_k, g_k, h_k)$ *with entries in* $\mathrm{N}(0,1)$. *For* $j = 1, 2, \ldots, k$, *let* $\xi_j = \|[S_j, g_j, h_j]\|_F$. *Orthonormalize the set of vectors* $\{\mathrm{vec}([S_j, g_j, h_j]) : j = 1, 2, \ldots, k\}$ *to get the set* $\{z_j \in \mathbb{R}^{n(n+2)} : j = 1, 2, \ldots, k\}$, *e.g., using QR factorization. Then for each* j, *let* $[S_j, g_j, h_j] = \xi_j \,\mathrm{unvec}(z_j)$.
2. *Let* $p = m(n+1)$. *Approximate* ω_k *and* ω_p *using* (5).
3. *For* $j = 1, 2, \ldots, k$, *calculate* $u_j = R^{-1}\left(g_j - S_j x + \|b - Ax\|_2 R^{-T} h_j\right)$. *Using the approximations for* ω_k *and* ω_p, *calculate the absolute condition vector*

$$\kappa_{\mathrm{abs}} = \frac{\omega_k}{\omega_p \sqrt{p}} \left| |u_1|^2 + |u_2|^2 + \cdots + |u_k|^2 \right|^{1/2}.$$

4. *Let the relative condition vector* κ_{rel} *be the vector* κ_{abs} *divided componentwise by* x, *leaving entries of* κ_{abs} *corresponding to zero entries of* x *unchanged.*

A similar method based on the averaged condition estimator of (8) can also be obtained. With such a method, the orthonormalization in Step 1 can be avoided.

Consider the SCE approach on Example 1. Algorithm 1 readily detects the differences in sensitivity of each of the components of the solution vector x. Recall that the first entry of the computed solution vector $\overline{x}$ is much more accurate than the second entry. Denote the relative condition vector obtained from applying the SCE method to data $[A, b]$ with k perturbations by $\kappa_{\mathrm{rel},k}([A, b])$. Then

$$\kappa_{\mathrm{rel},1}([A,b]) \approx \begin{bmatrix} 6 \times 10^7 \\ 1 \times 10^{24} \end{bmatrix},$$

$$\kappa_{\mathrm{rel},2}([A,b]) \approx \begin{bmatrix} 9 \times 10^7 \\ 2 \times 10^{24} \end{bmatrix},$$

$$\kappa_{\mathrm{rel},8}([A,b]) \approx \begin{bmatrix} 7 \times 10^7 \\ 1 \times 10^{24} \end{bmatrix}.$$

An averaged condition estimator-based method applied to the same data yields

$$\kappa_{\mathrm{rel},1}([A,b]) \approx \begin{bmatrix} 4 \times 10^7 \\ 8 \times 10^{23} \end{bmatrix},$$

$$\kappa_{\mathrm{rel},2}([A,b]) \approx \begin{bmatrix} 8 \times 10^7 \\ 2 \times 10^{24} \end{bmatrix},$$

$$\kappa_{\mathrm{rel},8}([A,b]) \approx \begin{bmatrix} 1 \times 10^8 \\ 2 \times 10^{24} \end{bmatrix}.$$

SCE clearly reveals the relative ill conditioning in the second component of the solution vector compared to the first component. We see this even in the $k = 1$ case. In fact, several perturbations are not necessary to get a good condition estimate; one or two perturbations are usually sufficient. However, note that the norm of the condition vector $\|\kappa_{\mathrm{rel},2}\|_2 \approx 10^{24}$ in this example is quite large compared to the Frobenius norm condition number $\kappa_F(A) \approx 10^8$. Each entry of $\kappa_{\mathrm{rel},2}$ is quite a bit larger than we would have expected based on our analysis of Example 1 in section 1. However, the discussion of section 1 assumed that errors were from rounding only, leading

to multiplicative rather than additive perturbations. Unlike additive perturbations, multiplicative perturbations have size relative to the size of the perturbed quantity. The next section describes how SCE can be adapted to handle such perturbations.

3. SCE for relative perturbations. Relative perturbations were introduced in section 1.3.1. The magnitudes of such perturbations are relative to the magnitudes of the corresponding entries in the input arguments (see (10)). These perturbations may arise from input error or from rounding error, and hence are the most common perturbations encountered in practice. In fact, LAPACK algorithms that provide componentwise error bounds [1, section 4.3.2] assume that perturbations of input data are relative. It is often the case that we wish to know the sensitivity of a function to relative perturbations of its arguments. For example, if errors in the input $[A, b]$ to the function

$$g([A, b]) = A^+ b$$

are due to rounding in a finite-word-length computer, then we would expect the mantissas of the computer-represented entries of $[A, b]$ to be perturbed by similar amounts, on the order of the computer's relative machine precision. This is true even though the exponents of the entries of $[A, b]$ may be quite different, leading to greatly differing absolute errors. SCE is flexible enough to accurately gauge the sensitivity of matrix functions subject to relative perturbations.

Let L be the function that multiplies each component of an $m \times (n+1)$ matrix, call it M, by the corresponding component of $[A, b] \in \mathbb{R}^{m\times(n+1)}$. The (i, j)th entry of the componentwise product is

$$L(M)_{ij} = [A, b]_{ij}\, m_{ij}.$$

Define $Z \in \mathbb{R}^{m\times(n+1)}$ to be the matrix of all ones. Then (compare (10) and (11))

$$L(Z + [E, f]) = L(Z) + L([E, f]) \tag{20}$$

$$= [A, b] + L([E, f]). \tag{21}$$

We see from (20)–(21) that L converts a general perturbation of Z into a relative perturbation of $[A, b]$. Therefore, to obtain the sensitivity of the solution with respect to relative perturbations, we simply evaluate the Fréchet derivative of

$$g([A, b]) = g(L(Z))$$

with respect to Z in the direction $[E, f]$, which is

$$\begin{aligned} \mathrm{D}g(Z; [E, f]) &= \mathrm{D}g(L(Z))\, \mathrm{D}L(Z; [E, f]) \\ &= \mathrm{D}g([A, b])\, L([E, f]) \\ &= \mathrm{D}g([A, b]; L([E, f])) \end{aligned}$$

since L is linear. Thus, to estimate the condition of the linear least squares solution when perturbations have relative magnitudes, we first generate the perturbations E and f and multiply them componentwise by the entries of A and b, respectively. Then we proceed as for general perturbations.

From section 1.2, the above discussion, and (16) we obtain an SCE-based method for estimating the condition of solutions to linear least squares problems with relative perturbations. Inputs to the method are $A \in \mathbb{R}_n^{m\times n}$ and $b \in \mathbb{R}^m$, and the output

is the relative condition vector κ_{rel}, an estimate of the relative sensitivity of each entry of the solution vector $x \in \mathbb{R}^n$. It is assumed that perturbations E_j and f_j have entrywise magnitudes relative to the magnitudes of A and b. The method requires the QR factors of A that are assumed to have been calculated during the solution of the linear least squares problem. The integer $k \geq 1$ refers to the number of perturbations of input data ($k = 1$ or 2 generally being adequate in practice).

ALGORITHM 2 (subspace condition estimation for relative perturbations).

1. *Generate* $(E_1, f_1), (E_2, f_2), \ldots, (E_k, f_k)$ *with entries in* $\mathrm{N}(0,1)$ *and orthonormalize them* ($E_j \in \mathbb{R}^{m \times n}$, $f_j \in \mathbb{R}^m$).
2. *For* $j = 1, 2, \ldots, k$, *set* $[E_j, f_j]$ *equal to the componentwise product of* $[A, b]$ *with* $[E_j, f_j]$.
3. *Let* $p = m(n+1)$. *Approximate* ω_k *and* ω_p *using* (5).
4. *For* $j = 1, 2, \ldots, k$, *calculate* $u_j = R^{-1}\left(Q_1(f_j - E_j x) + R^{-T}E_j^T(b - Ax)\right)$. *Using the approximations for* ω_p *and* ω_k, *calculate the absolute condition vector*
$$\kappa_{\text{abs}} = \frac{\omega_k}{\omega_p}\left||u_1|^2 + |u_2|^2 + \cdots + |u_k|^2\right|^{1/2}.$$
5. *Let the relative condition vector* κ_{rel} *be the vector* κ_{abs} *divided componentwise by* x, *leaving entries of* κ_{abs} *corresponding to zero entries of* x *unchanged.*

Applying Algorithm 2 to A and b of Example 1, we find that

$$\kappa_{\text{rel},1}([A,b]) \approx \begin{bmatrix} 2 \\ 7 \times 10^{15} \end{bmatrix},$$
$$\kappa_{\text{rel},2}([A,b]) \approx \begin{bmatrix} 0.2 \\ 1 \times 10^{16} \end{bmatrix},$$
$$\kappa_{\text{rel},9}([A,b]) \approx \begin{bmatrix} 1 \\ 2 \times 10^{16} \end{bmatrix}.$$

These results agree with the discussion of Example 1 in section 1.

Applying Algorithm 2 to A and b of Example 1, we find that

$$\kappa_{\text{rel},1}([A,b]) \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$
$$\kappa_{\text{rel},2}([A,b]) \approx \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$
$$\kappa_{\text{rel},9}([A,b]) \approx \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$$

SCE clearly reveals the insensitivity of the linear least squares solution of Example 1 to relative perturbations when b is well aligned with $\mathcal{R}(A)$. However, applying Algorithm 2 to A and c of Example 2, we find that

$$\kappa_{\text{rel},1}([A,c]) \approx \begin{bmatrix} 2 \times 10^{16} \\ 8 \times 10^{15} \end{bmatrix},$$
$$\kappa_{\text{rel},2}([A,c]) \approx \begin{bmatrix} 6 \times 10^{15} \\ 3 \times 10^{15} \end{bmatrix},$$
$$\kappa_{\text{rel},9}([A,c]) \approx \begin{bmatrix} 1 \times 10^{16} \\ 6 \times 10^{15} \end{bmatrix}.$$

In this case c is not well aligned with $\mathcal{R}(A)$, and the SCE method reveals that the entrywise condition numbers of this least squares problem are on the order of $\kappa_2^2(A)$.

4. SCE for structured perturbations. Perturbations of the input data to a linear least squares problem may be restricted to a specific structure. For example, they may be constrained by physical properties of the system being modeled or by the mathematical form of the modeling matrices. Perturbation structure may also arise as a consequence of the method used to obtain the solution. For example, the solution of an upper triangular system by backsolving does not involve the lower triangular zero entries of the data matrix; hence, perturbing these entries is not necessary, and may lead to an inaccurate sensitivity estimate.

Frequently, we wish to estimate the sensitivity of matrix functions when only a restricted class of perturbations is allowed (see [13, section 2.1] for several references). For example (see section 6.1), in linear regression one column of the data matrix may correspond to a constant term and is not subject to perturbations. If our data matrix is the incidence matrix for an electrical circuit, perturbations will not affect the numerous zero entries of the matrix, and the nonzero entries will be perturbed symmetrically. SCE can handle these and other forms of perturbation easily.

Suppose the augmented matrix $[A, b]$ of a linear least squares problem is generated from a vector $z \in \mathbb{R}^p$ by an affine function $F : \mathbb{R}^p \to \mathbb{R}^{m\times(n+1)}$

$$F(z) = L(z) + K = [A, b],$$

where L is a linear function and K is a constant matrix. We can express the linear least squares function as

$$g([A, b]) = g(F(z)). \tag{22}$$

For example, suppose our linear least squares data consist of a truncated symmetric Toeplitz matrix $T \in \mathbb{R}^{4\times3}$ and a vector $v \in \mathbb{R}^4$, whose first two entries we assume will not be perturbed:

$$[T|v] = \left[\begin{array}{ccc|c} t_1 & t_2 & t_3 & k_1 \\ t_2 & t_1 & t_2 & k_2 \\ t_3 & t_2 & t_1 & v_3 \\ t_4 & t_3 & t_2 & v_4 \end{array}\right].$$

Then we can represent the data by the functions

$$F(z) = \begin{bmatrix} z_1 & z_2 & z_3 & k_1 \\ z_2 & z_1 & z_2 & k_2 \\ z_3 & z_2 & z_1 & z_5 \\ z_4 & z_3 & z_2 & z_6 \end{bmatrix}$$

$$= L(z) + K = \begin{bmatrix} z_1 & z_2 & z_3 & 0 \\ z_2 & z_1 & z_2 & 0 \\ z_3 & z_2 & z_1 & z_5 \\ z_4 & z_3 & z_2 & z_6 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & k_1 \\ 0 & 0 & 0 & k_2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

evaluated at the point

$$z_{[T,v]} = [t_1, t_2, t_3, t_4, v_3, v_4]^T.$$

The Fréchet derivative of (22) with respect to z evaluated in the direction $\tilde{z}$ is

$$\begin{aligned} \mathrm{D}g(z; \tilde{z}) &= \mathrm{D}g(F(z))\, \mathrm{D}F(z; \tilde{z}) \\ &= \mathrm{D}g([A, b])\, L(\tilde{z}) \\ &= \mathrm{D}g([A, b]; L(\tilde{z})) \end{aligned}$$

since $F = L + K$ is affine. The SCE method respects the perturbation structure by perturbing only the generating vector z by $\tilde{z}$, and then linearly mapping $\tilde{z}$ to the correct matrix and vector shapes corresponding to structured perturbations E and f.

From section 1.2, the above discussion, and (16) we obtain an SCE-based method for estimating the condition of solutions to linear least squares problems with structured perturbations. Inputs to this method are $A \in \mathbb{R}_n^{m\times n}$ and $b \in \mathbb{R}^m$, where $[A, b]$ has been obtained from an affine function $F = L + K$. The method outputs the relative condition vector κ_{rel}, an estimate of the relative sensitivity of each entry of the solution vector $x \in \mathbb{R}^n$. It is assumed that perturbations $z_j \in \mathbb{R}^p$ are constrained by the linear structure map L. The method requires the QR factors of A that are assumed to have been previously calculated during the solution of the linear least squares problem. The random vectors $z_j \in \mathbb{R}^p$ are mapped by L to the structured perturbations $E_j \in \mathbb{R}^{m\times n}$ and $f_j \in \mathbb{R}^m$. The integer $k \geq 1$ refers to the number of perturbations of input data ($k = 1$ or 2 generally being adequate in practice).

ALGORITHM 3 (subspace condition estimation for structured perturbations).

1. *Generate* $z_1, z_2, \ldots, z_k$ *with entries in* $\mathrm{N}(0, 1)$ *and orthonormalize them* ($z_j \in \mathbb{R}^p$).
2. *Approximate* ω_k *and* ω_p *using* (5).
3. *For* $j = 1, 2, \ldots, k$, *set* $[E_j, f_j] = L(z_j)$.
4. *For* $j = 1, 2, \ldots, k$, *calculate* $u_j = R^{-1}\left(Q_1(f_j - E_j x) + R^{-T}E_j^T(b - Ax)\right)$. *Using the approximations for* ω_k *and* ω_p, *calculate the absolute condition vector*

$$\kappa_{\text{abs}} = \frac{\omega_k}{\omega_p}\left||u_1|^2 + |u_2|^2 + \cdots + |u_k|^2\right|^{1/2}.$$

5. *Let the relative condition vector* κ_{rel} *be the vector* κ_{abs} *divided componentwise by* x, *leaving entries of* κ_{abs} *corresponding to zero entries of* x *unchanged.*

When we apply Algorithm 3 to Example 3, we find that

$$\kappa_{\text{rel},1} \approx \begin{bmatrix} 0.7 \\ 0.9 \end{bmatrix},$$

$$\kappa_{\text{rel},2} \approx \begin{bmatrix} 0.9 \\ 3 \end{bmatrix},$$

$$\kappa_{\text{rel},6} \approx \begin{bmatrix} 0.8 \\ 2 \end{bmatrix}.$$

SCE is able to respect the structure of the perturbations in this example, providing a more realistic estimate of the sensitivity than the 2-norm condition number $\kappa_2(A) \approx 10^8$. Example 3 might arise in a linear regression situation in which the sample times are considered to be exact, for instance.

5. Comparison with existing algorithms. SCE applied to the full column rank linear least squares problem requires about the same amount of computation as other condition estimation methods. A MATLAB implementation of Algorithm 1 written by the authors requires $\mathrm{O}(n^2k^2 + m)$ flops, where the data matrix is $m \times n$ and k is the number of perturbations of the input data. An averaged condition estimator-based algorithm requires $\mathrm{O}(n^2k + m)$ flops. For most problems, $k = 1$ is sufficient. The m term represents the cost of calculating the norm of the residual. This quantity can be calculated during QR factorization. Standard methods that estimate the sensitivity of the solution to the least squares problem [3], [8], [10], [18],

[25] require $O(n^2)$ flops after QR factorization; however, for these methods to obtain results as accurate as those provided by the SCE method, they also must calculate the norm of the residual. If we know a priori that the 2-norm of b is equal to a constant β and we accumulate Householder transformations on b during QR factorization (giving us Qb), then we can calculate the norm of the residual in $O(n)$ flops (see (15)) by

$$\|b - Ax\|_2 = \sqrt{\beta^2 - \|Q_1 b\|_2^2}.$$

The additional work required to obtain a condition estimate of the linear least squares problem is small compared to the work to solve the linear least squares problem itself, which is on the order of mn^2 flops if done by QR factorization. The ratios of flops required to obtain condition estimates to flops required to solve the linear least squares problem are

$$\begin{aligned} \frac{k^2}{m} &\approx \frac{1}{m} \text{ for subspace SCE,} \\ \frac{k}{m} &\approx \frac{1}{m} \text{ for averaged SCE,} \\ &\quad \frac{1}{m} \text{ for standard methods.} \end{aligned}$$

In the case of more general full-column-rank linear least squares problems of the form

$$\min_X \|AX - B\|_2, \tag{23}$$

where $A \in \mathbb{R}_n^{m\times n}$ and $B \in \mathbb{R}^{m\times p}$, SCE algorithms can obtain condition estimates in approximately order n^2p flops, while the solution of the linear least squares problem requires order mn^2p flops; hence, the condition-estimate-to-solution flop ratio is again about $1/m$. Condition estimation of (23) is easily done by solving p independent standard linear least squares problems, each corresponding to one column of X and one column of B. The QR factors are obtained once and then reused for each column of the solution and of the condition estimate.

6. Examples. The authors have applied SCE to examples from various sources. Some of these examples are discussed below.

6.1. Linear regression. Suppose we collect data with a system that provides a sample every microsecond t_i starting at $t_1 = 10^{-6}$. Assume that we have obtained the data $b_i = 12, 13, 18, 27$ for $i = 1, 2, 3, 4$, respectively, which we believe have an approximately affine relationship; that is, we want to determine the line $b = \alpha t + \beta$ that most closely fits the data, where b is the vector comprised of the data samples b_i. Define

$$A = \begin{bmatrix} t_1 & 1 \\ t_2 & 1 \\ t_3 & 1 \\ t_4 & 1 \end{bmatrix} \qquad \text{and} \qquad x = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}.$$

We want the minimum 2-norm solution to the linear least squares problem $\min_x \|Ax - b\|$. The 2-norm condition number of A is $\kappa_2(A) \approx 10^6$; however, MATLAB gives the exact solution

$$x = \begin{bmatrix} 5 \times 10^6 \\ 5 \end{bmatrix}.$$

Realistically, we would expect no perturbation in the second column of A and only small relative perturbations in the first column of A. Applying SCE for structured and relative perturbations to $[A, b]$, we get the relative condition vector

$$\kappa_{\text{rel}}([A, b]) \approx \begin{bmatrix} 2 \\ 4 \end{bmatrix},$$

indicating that this linear least squares problem is well conditioned despite a large $\kappa_2(A)$.

6.2. 8×8 Vandermonde. This example is a generalization of one from [13, section 2]. Let $A \in \mathbb{R}^{8\times 8}$ have entries $a_{ij} = j^{2(i-1)}$, let $b = e_1 \in \mathbb{R}^8$, and let c equal the sum of the columns of A. Now $\kappa_2(A) \approx 10^{13}$, but Algorithm 2 yields condition estimates $\kappa_{\text{rel}}([A, b])$ and $\kappa_{\text{rel}}([A, c])$, whose largest entries are on the order of 10^5 and 10^8, respectively. In his discussion of the example in [13, section 2], Higham observes that a normwise error bound can grossly overestimate the true sensitivity of a solution when perturbations are restricted to be of a certain type. The example clearly illustrates this.

6.3. Longley. The well-known Longley regression problem [2], [19] is quite ill conditioned, with $\kappa_2(\text{Longley}) \approx 10^9$. Nevertheless, componentwise relative condition estimation can tighten the upper bound on the sensitivity of the solution, as was observed in [13, section 4]. Algorithm 2 yields the condition vector

$$\kappa_{\text{rel}}(\text{Longley}) \approx 10^3 \begin{bmatrix} 5 & 90 & 20 & 5 & 4 & 80 & 5 \end{bmatrix}^T$$

for the Longley data. The last column of the Longley data matrix lists years in which the row of data was obtained. Presumably, this column is not subject to input error, nor is the first column, which corresponds to the linear regression constant. If we use a combination of Algorithms 2 and 3 to reflect this perturbation structure, we obtain an even tighter bound on componentwise condition

$$\kappa_{\text{rel}}(\text{Longley}) \approx 10 \begin{bmatrix} 7 & 200 & 30 & 8 & 5 & 200 & 7 \end{bmatrix}^T.$$

7. Conclusion. The general theory of SCE, a statistically based method of obtaining componentwise sensitivity for general matrix functions, is described in [16]. The method is based on perturbing the inputs to matrix functions and measuring the resulting changes in the outputs. The application of SCE to nonsingular linear systems is examined in [17]. In this paper we have applied SCE to the full-column-rank linear least squares problem $\min_x \|Ax - b\|_2$, where $A \in \mathbb{R}_n^{m\times n}$ and $b \in \mathbb{R}^m$. Here, SCE takes advantage of the availability of the explicit Fréchet derivative of the linear least squares problem. Furthermore, taking advantage of the QR factors obtained in the primary function evaluation allows us to avoid doing additional full function evaluations, thus keeping computational costs for SCE to a level comparable with conventional condition estimators. Moreover, SCE has several advantages over other methods. Most importantly, it provides a full vector of condition numbers rather than a single number, as in norm-based condition estimation approaches. Hence, it has the ability to reveal component-specific sensitivities in the solution to a linear least squares problem. SCE also considers all input data when forming its condition estimates, not merely the data matrix A. Thus, more reliable results can be obtained reflecting the relative alignment of the observation vector b with respect to the range of A.

The flexibility of SCE allows it to be adapted to problems with perturbations of small relative magnitudes and perturbations obeying specific structural constraints, thus providing condition numbers that more realistically measure the sensitivity of such problems. A rigorous statistical theory exists for SCE that describes how likely its condition estimate lies within a specified range of the true condition. The probability of accuracy can be increased by doing additional perturbations of the input arguments. Software for the SCE method, written in MATLAB by the authors, has performed successfully on a wide variety of linear least squares problems.

Acknowledgments. The authors thank Shivkumar Chandrasekaran, Thorkell Gudmundsson, and two anonymous reviewers for providing several helpful suggestions during the preparation of this paper.

REFERENCES

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen, LAPACK *Users' Guide*, 2nd ed., SIAM, Philadelphia, PA, 1994.

[2] A. E. Beaton, D. B. Rubin, and J. L. Barone, *The acceptability of regression solutions: Another look at computational accuracy*, J. Amer. Statist. Assoc., 71 (1976), pp. 158–168.

[3] Å. Björck, *Least squares methods*, in Handbook of Numerical Analysis, Vol. 1, P. G. Ciarlet and J. L. Lions, eds., Elsevier–North-Holland, New York, 1992, pp. 465–652.

[4] S. Chandrasekaran, *When Is a Linear System Ill-Conditioned?*, Ph.D. thesis, Yale University, New Haven, CT, 1994.

[5] S. Chandrasekaran and I. C. F. Ipsen, *On the sensitivity of solution components in linear systems of equations*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 93–112.

[6] A. K. Cline, C. B. Moler, G. W. Stewart, and J. H. Wilkinson, *An estimate for the condition number of a matrix*, SIAM J. Numer. Anal., 16 (1979), pp. 368–375.

[7] M. H. DeGroot, *Probability and Statistics*, 2nd ed., Addison–Wesley, Reading, MA, 1989.

[8] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, LINPACK *Users' Guide*, SIAM, Philadelphia, PA, 1979.

[9] 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.

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

[11] G. H. Golub and J. H. Wilkinson, *Note on the iterative refinement of least squares solution*, Numer. Math., 9 (1966), pp. 139–148.

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

[13] N. J. Higham, *A survey of componentwise perturbation theory in numerical linear algebra*, in Proc. Symposia in Applied Mathematics, Mathematics of Computation 1943–1993: A Half Century of Computational Mathematics 48, Walter Gautschi, ed., AMS, Providence, RI, 1994, pp. 49–77.

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

[15] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 2nd ed., Prentice–Hall, Englewood Cliffs, NJ, 1988.

[16] C. S. Kenney and A. J. Laub, *Small-sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comput., 15 (1994), pp. 36–61.

[17] C. S. Kenney, A. J. Laub, and M. S. Reese, *Statistical condition estimation for linear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 566–583.

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

[19] J. W. Longley, *An appraisal of least squares programs for the electronic computer from the point of view of the user*, J. Amer. Statist. Assoc., 62 (1967), pp. 819–841.

[20] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw–Hill, New York, 1991.

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

[22] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.
[23] G. W. Stewart, *Perturbation theory and least squares with errors in the variables*, in Proc. AMS Workshop on Measurement Error Models, Humboldt State University, Arcata, CA, 1989, pp. 171–181.
[24] G. W. Stewart, *Stochastic perturbation theory*, SIAM Rev., 32 (1990), pp. 579–610.
[25] G. W. Stewart and J. Sun, *Matrix Perturbation Theory*, Academic Press, San Diego, CA, 1990.
[26] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, PA, 1991.
[27] P.-Å. Wedin, *Perturbation theory for pseudo-inverses*, BIT, 13 (1973), pp. 217–232.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 924–932, October 1998

007

GLOBAL BLOCK-SIMILARITY AND POLE ASSIGNMENT OF CLASS C^{P*}

JOSEP FERRER† AND FERRAN PUERTA†

Abstract. Starting from the existence of a C^p-basis for any C^p-family of subspaces having constant dimension, we construct a Brunovsky basis of class C^p for a C^p-family of pairs of matrices having constant Brunovsky type. We derive a global pole assignment theorem for such kinds of pairs. In all the cases we assume that the manifold of parameters is contractible.

Key words. block-similarity, Brunovsky, pole assignment, principal bundles, families of subspaces

AMS subject classifications. 93B, 53C

PII. S0895479893244018

1. Introduction.

There is an abundance of literature concerning parametrized families of linear systems. "Most results attempt in essence to establish local-global principles: does pointwise solvability imply the existence of nicely parametrized solution?" [17].

In particular, there is also a wealth of literature concerning the problems dealt with in this paper—global reduction to the canonical form and global pole assignment. Both problems have been studied widely for pointwise controllable families. One of the aims of this paper is to present a generalization to families non-necessarily pointwise controllable.

For a general introduction to families of linear systems, see, for example, [13]. Some problems that justify the study of families of systems are presented there, and one tackles the classification of families (fine moduli spaces), the existence of global canonical forms, and some others. In fact, [13] deals with a more general class of families of linear systems in terms of bundles over the space of parameters, which includes the parametrized ones.

An alternative generalization is the consideration of systems over rings—this is to say, pairs of matrices with entries on a commutative ring. The particular case of parametrized families arises when rings of functions defined in the space of parameters are considered. See [15] and [2] for a general introduction, and [17] for a survey and many references mainly centered on control and stabilization problems.

The Swan theorem connects these approaches to local-global problems by means of the correspondence between vector bundles and projective modules.

Let us now see in more detail the problems explicitly studied in this paper. The central one is Theorem 4.1 in section 5: the existence of a differentiable reduction to the Brunovsky form (or equivalently, the existence of a differentiable Brunovsky basis) for a differentiable family of pairs of matrices having constant Brunovsky type, provided that the manifold of parameters is contractible.

For pointwise completely controllable families, the result is classical in the continuous case [2] or for polynomial rings [18]. It is also well known in the differentiable

*Received by the editors February 8, 1993; accepted for publication (in revised form) by S. Van Huffel July 9, 1997; published electronically June 9, 1998.
http://www.siam.org/journals/simax/19-4/24401.html
†Departament de Matemàtica Aplicada I, ETSEIB-UPC Avda., Diagonal 647, 08028 Barcelona, Spain (ferrer@ma1.upc.es, puerta@ma1.upc.es).

case and used in the literature without explicit reference (see, for example, [14], [17]). On the other hand, the global reduction to the Jordan form for a family of square matrices has been more recently studied by [4], [8], and [5].

However, the extension to the general case starting on the above "extreme" cases (controllable, Jordan) does not seem trivial in spite of the global splitting in [13, Section 11.3.1] because this is not a direct split.

In fact, this obstruction is not surprising because it appears in other problems concerning parametrized families of systems. For example, the construction of a versal deformation has been solved in [1] for square matrices and in [18] for pointwise controllable pairs of matrices. But the general case does not derive from them (see [6]).

Second, we study the global pole assignment of parametrized families, which has been widely studied (see, for example, [17]), always under the hypothesis of pointwise controllability. The problem is solved in [2] and [16] by means of algebro-geometric and algebraic techniques, respectively, if provided constant controllability indices. Other conditions are considered, for example, in [19] (constant rank of B, ring controllability) or in [12] (one-dimensional manifold of parameters).

Here the general result for non-necessarily pointwise controllable families (Theorem 5.1, in section 5) is derived as an application of the previous one about global reduction to Brunovsky form. By means of it, the proof in [9] of pole assignment for a pair of matrices (non-necessarily controllable) can be immediately translated into a differentiable family of pairs of matrices having constant Brunovsky type.

Going back to the central result about global reduction to the Brunovsky form, our technique consists essentially of extending the construction in the constant case to the parametrized one by means of Theorem 2.2, in section 2—the existence of global differentiable bases for a differentiable family of subspaces having constant dimension, parametrized over a contractible manifold.

A classical and fundamental reference for this basic tool in the continuous case is [10]. In fact, it proves by means of the method of cocycles a generalization for operator valued functions defined on a contractible manifold, provided it is compact. In our case of matrix valued differentiable functions on a contractible manifold non-necessarily compact, it is used, for example, in [13], and it is explicitly presented in [20] (remark after Section 2.5).

For the continuous case, the key point is [11, Section 3.4.8] about the triviality of bundles over contractible spaces. Then, the smooth case follows by means of approximation theorems. Alternatively, we have checked that the proof of this key point in [11] for the continuous case can be adapted to the differentiable one.

Notice that the hypothesis that the manifold of parameters must be contractible is only used for Proposition A to be verified. Then, all the machinery works under other conditions whenever Proposition A holds (for example, under the conditions of [11, Section 8.1.2]).

This technique for extending the local case to the parametrized one is possible if we have a geometrical description of the construction in the constant case (in terms of kernels, supplementary subspaces, ...). In this case, we use the description of a Brunovsky basis obtained in [7] as a basis of the global space adapted to an increasing chain of certain subspaces.

The minimum subspace of this chain corresponds to the uncontrollable subsystem, whose Brunovsky bases are, in fact, Jordan bases. For this step, we remark that in [5] the same technique has been used for the case of square matrices having constant

Jordan type, by means of the usual description of Jordan bases as adapted to an increasing chain of kernels.

Notice that this method does not need further references to algebro-geometric or algebraic techniques, but only Proposition A and the standard machinery for the constant case. As one of the referees has pointed out, this presentation seems more accessible to engineers.

Moreover, our approach has some connections with that in [3]. There, the control properties of a pair of matrices (A, B) are studied by means of the geometric properties of the curve $\gamma : \mathbb{R} \longrightarrow Gr_{s,n}$, $\gamma(\tau) = \text{Im } (e^{A\tau}B)$. Here, we do not consider a pair (A, B), but a C^p-family of pairs $(A(t), B(t))$, $t \in M$, having constant Brunovsky type. Then, instead of a curve in $Gr_{s,n}$, we could consider a multiparametrized mapping $\Gamma : M \times \mathbb{R} \longrightarrow Gr_{s,n}$, $\Gamma(t, \tau) = \text{Im } (E^{A(t)\tau}B(\tau))$. It can be expected that some control properties of the family should be related to the geometric properties of Γ.

In sections 2 and 3 we recall, respectively, the basic facts about differentiable families of subspaces and about global similarity of square matrices which we will use in the sequel. See [5] for more details. Section 4 is devoted to the proof of the main Theorem 4.1. And section 5 contains the application to global pole assignment.

Throughout the paper K denotes $\mathbb{R}$ or $\mathbb{C}$, $(e_1, \ldots, e_n)$ the standard basis of K^n, and $Gr_{k,n}$ the set of k-dimensional subspaces of K^n. If $v_1, \ldots, v_s$ are vectors of K^n, then $[v_1, \ldots, v_s]$ will denote the subspace spanned by them.

We write $M_{n\times k}(K)$ for the vector space of $(n \times k)$-matrices with entries in K, $M^*_{n\times k}$ the open subset formed by the matrices $A \in M_{n\times k}(K)$ having rank k $(\leq n)$, and $Gl(n)$ the linear group of nonsingular matrices of $M_n(K)$. If $A \in M_{n\times m}(K)$, we also denote by A the linear map from K^m to K^n defined by $(x_1, \ldots, x_m) \longrightarrow (x_1, \ldots, x_m)A^t$, where A^t is the transpose matrix of A. *Id* will denote the identity mapping, and Id_k the identity k-matrix.

By a differentiable manifold we mean a C^p-manifold, $1 \leq p \leq \infty$. Throughout the paper M will be a differentiable manifold. In the same way, by a differentiable map between two of such manifolds, we mean a C^p-morphism.

2. Differentiable families of subspaces. In $Gr_{k,n}$ the usual topology and differentiable structure is considered. With them, $Gr_{k,n}$ is a compact homogeneous manifold. This topology is equivalent to the one induced by the *gap metric* [9]. A proof of this equivalence is given in [5].

By a *family of k-subspaces* of K^n parametrized on M we mean a map $\mathcal{L} : M \longrightarrow Gr_{k,n}$. If it is differentiable, we write $\mathcal{L} \in C^p(M, Gr_{k,n})$. The existence of "differentiable local basis" is a useful criterium for the differentiability of a family of subspaces. It follows immediately from the local triviality of a bundle.

PROPOSITION 2.1. *Let M be a differentiable manifold, and $\mathcal{L} : M \longrightarrow Gr_{k,n}$ a family of k-subspaces parametrized on M. Then, $\mathcal{L}$ is differentiable if and only if: for every $t_0 \in M$ there is an open neighborhood W_{t_0} of t_0 in M, and k maps $v_i \in C^p(W_{t_0}, K^n)$, $1 \leq i \leq k$, such that $\{v_1(t), \ldots, v_k(t)\}$ is a basis of $\mathcal{L}(t)$, for all $t \in W_{t_0}$.*

For example, if $\mathcal{L} \in C^p(M, Gr_{k,n})$, then $\mathcal{L}^\perp \in C^p(M, Gr_{n-k,n})$, where $\mathcal{L}^\perp(t) = \mathcal{L}(t)^\perp$. Also, if $A \in C^p(M, M_{n\times m}(K))$, with rank $A(t) = k$, for all $t \in M$, then Im $A \in C^p(M, Gr_{k,n})$ and Ker $A \in C^p(M, Gr_{n-k,n})$.

Let us assume that M is contractible. As we have remarked in the introduction, a basic tool in our technique is the existence of a "differentiable global basis"(in fact, we will use the corollary).

THEOREM 2.2 (see the introduction for the references). *Let M be a contractible*

manifold, and $\mathcal{L} \in C^p(M, Gr_{k,n})$ *a differentiable family of* k*-subspaces parametrized on* M*. Then there exist* k *maps* $v_i \in C^p(M, K^n)$, $1 \le i \le k$, *such that* $\{v_1(t), \dots, v_k(t)\}$ *is a basis of* $\mathcal{L}(t)$ *for every* $t \in M$*.*

COROLLARY 2.3. *Let* M *be a contractible manifold, and* $\mathcal{L}_i \in C^p(M, Gr_{k_i,n})$, $1 \le i \le 2$, *such that* $\mathcal{L}_1(t) \subset \mathcal{L}_2(t)$, *for all* $t \in M$*. Then there exist* $v_1, \dots, v_k \in C^p(M, K^n)$, $k = k_2 - k_1$, *such that*

$$\mathcal{L}_2(t) = \mathcal{L}_1(t) \oplus [v_1(t), \dots, v_k(t)], \quad \text{for all} \quad t \in M.$$

3. Global similarity of class C^p. As we have said in the introduction, the first step in the proof of Theorem 4.1 about block-similarity of class C^p is the already known analogous result concerning similarity. Let M be a differentiable manifold, and $A \in C^p(M, M_n(\mathbb{C}))$; that is to say, A$(t)$ is a family of n-square complex matrices, parametrized by $t \in M$, and of class C^p. The family $A(t)$ is said to have *constant Jordan type* if the number of distinct eigenvalues and the list of the sizes of the Jordan blocks corresponding to different eigenvalues are independent of t.

PROPOSITION 3.1. *Let* M *be a simply connected manifold, and* $A \in C^p(M, M_n(\mathbb{C}))$ *having constant Jordan type. Then:*

(i) *there exist* $\lambda_1, \dots, \lambda_q \in C^p(M, \mathbb{C})$ *such that* $\lambda_1(t), \dots, \lambda_q(t)$ *are the* q *distinct eigenvalues of* $A(t)$*, for every* $t \in M$*;*

(ii) *the respective algebraic multiplicities* $m_1, \dots, m_q$ *of these eigenvalues are constant.*

THEOREM 3.2. *Let* M *be a contractible manifold, and* $A \in C^p(M, M_n(\mathbb{C}))$ *having constant Jordan type. Then, there exists* $S \in C^p(M, Gl(n))$ *such that* $S(t)^{-1}A(t)S(t)$ *is a Jordan matrix, for all* $t \in M$*.*

4. Global block-similarity of class C^p.

4.1. Brunovsky form. We recall some basic properties of the block-similarity of pairs of matrices. Let us consider pairs of matrices $(A \quad B)$, where $A \in M_n(\mathbb{C})$ and $B \in M_{n\times m}(\mathbb{C})$. Two of such pairs $(A \quad B)$ and $(A' \quad B')$ are called *block-similar* if there are complex matrices $S \in Gl(n)$, $T \in Gl(m)$, and $C \in M_{m\times n}$ such that $A' = S^{-1}(A + BCS^{-1})S$, $B' = S^{-1}BT$, or, equivalently,

$$(A' \quad B') = S^{-1}(A \quad B)\begin{pmatrix} S & 0 \\ C & T \end{pmatrix}.$$

Every pair $(A \quad B)$ is block-similar to its so-called *Brunovsky form*:

$$\begin{pmatrix} N_1 & & & & E_1 & & & \\ & N_2 & & & & E_2 & & \\ & & \cdots & & & & \cdots & \\ & & & N_r & & & & E_r \\ & & & & J & & & 0 \end{pmatrix},$$

where $k_1 \ge \cdots \ge k_r$, N_i is a nilpotent k_i-matrix, E_i is the column $(k_i \times 1)$-matrix transpose of $(0 \dots 0 \quad 1)$, and J is a Jordan matrix. This canonical form is unique, up to permutations of the Jordan blocks in J. In particular, $r = \text{rank}B$, and $s \equiv k_1 + \ldots + k_r = \text{rank}\ (B\ AB\ \ldots\ A^{n-1}B)$.

Numbers $k_1, \dots, k_r$ are called the *controllability indices* of $(A \quad B)$. Although the Jordan matrix J is not uniquely determined, its similarity invariants are well defined; we will refer to them as the *Jordan invariants of the pair* $(A \quad B)$. In particular, the *eigenvalues of the pair* $(A \quad B)$ are those of J.

Two pairs are block-similar if and only if they have the same Brunovsky form. Thus, the controllability indices and the Jordan invariants of the pair form a complete family of invariants for the block-similarity.

Analogous considerations are valid for pairs of matrices of the form $\binom{A}{B}$. By duality, both constructions are equivalent. In particular, the controllability indices and the Jordan invariant of a pair $(A \quad B)$ coincide with those of its transpose $\binom{A^t}{B^t}$.

4.2. Brunovsky bases. We shall follow the method in [7] for the construction of Brunovsky bases. We recall that, up to reordering, a Brunovsky basis of a pair $\binom{A}{B}$ is obtained there by successive extensions in the chain of subspaces

$$Y_n = \cdots = Y_N \subset Y_{N-1} \subset Y_{N-2} \subset \cdots \subset Y_2 \subset Y_1 \subset Y_0 \equiv \mathbb{C}^n \subset Y_{-1} \equiv \mathbb{C}^{n+m}$$

(the inclusions are strict) where

$$Y_i = \operatorname{Ker} \begin{pmatrix} B \\ BA \\ \cdots \\ BA^{i-1} \end{pmatrix}, \quad 1 \leq i \leq n\,.$$

One verifies that $A(Y_N) \subset Y_N$. And for $0 \leq i \leq N-1$, if $\overline{Y}_i$ is any complementary subspace to Y_{i+1} in Y_i, this is to say $Y_i = Y_{i+1} \oplus \overline{Y}_i$, then: A is injective on $\overline{Y}_i$; $A(\overline{Y}_i) \subset Y_{i-1}$; $A(\overline{Y}_i) \cap Y_i = \{0\}$.

Moreover, in this description the controllability indices are characterized as the conjugate partition of the one formed by the differences $\dim Y_i - \dim Y_{i+1}$, and the Jordan invariants of the pair $\binom{A}{B}$ are those of the endomorphism $A| : Y_N \longrightarrow Y_N$, defined as the restriction of A. In particular, $k_1 = N$, $r = n - \dim Y_1$, and $s = n - \dim Y_N$.

4.3. Constant Brunovsky type. Now let us consider families of pairs of matrices. Let M be a differentiable manifold, and $(A \quad B) \in C^p(M, M_{n\times(n+m)}(\mathbb{C}))$, that is to say, $(A(t) \quad B(t))$ is a family of pairs of matrices, where $A \in C^p(M, M_n(\mathbb{C}))$ and $B \in C^p(M, M_{n\times m}(\mathbb{C}))$. We say that the family $(A(t) \quad B(t))$ has *constant Brunovsky type*, if

(i) the controllability indices are constant,
(ii) the Jordan invariants have constant type.

Notice that, with regard to the description in section 4.2, it is equivalent to saying that:

(i′) the dimensions of the subspaces $Y_i(t)$ are independent of t,
(ii′) the family of endomorphisms $A(t)| : Y_N(t) \longrightarrow Y_N(t)$ has constant Jordan type.

4.4. Global Brunovsky bases. For such a differentiable family $(A(t) \quad B(t))$ having constant Brunovsky type, the question lies in the construction of a Brunovsky basis depending differentiably on $t \in M$, provided that M is contractible.

THEOREM 4.1. *Let M be a contractible manifold, and $(A \;\; B) \in C^p(M, M_{n\times(n+m)}(\mathbb{C}))$ a differentiable family of pairs of matrices having constant Brunovsky type. Then, there exist $S \in C^p(M, Gl(n))$, $T \in C^p(M, Gl(m))$, and $C \in C^p(M, M_{m\times n}(\mathbb{C}))$ such that*

$$S(t)^{-1}(A(t) \quad B(t)) \begin{pmatrix} S(t) & 0 \\ C(t) & T(t) \end{pmatrix}$$

is a Brunovsky matrix, for all $t \in M$.

Proof. Our aim is to adapt to parametrized pairs the construction relative to a constant pair $(A \quad B)$ sketched in section 4.2. In fact, as it was the case there, we shall deal with pairs of the form $\binom{A(t)}{B(t)}$. In order to do so, we shall consider the chain of subspaces

$$Y_n(t) = \cdots = Y_N(t) \subset Y_{N-1}(t) \subset \cdots \subset Y_1(t) \subset Y_0(t) \equiv \mathbb{C}^n \subset \mathbb{C}^{n+m},$$

where

$$Y_i(t) = \mathrm{Ker} \begin{pmatrix} B(t) \\ B(t)A(t) \\ \cdots \\ B(t)A(t)^{i-1} \end{pmatrix}, \quad 1 \leq i \leq n.$$

Because of hypothesis (i′) and the examples after Proposition 2.1, each term in the chain is a differentiable family of subspaces having constant dimension. The desired basis shall be constructed by successive extensions in this chain. The key point is the application of Theorem 3.2 in the first step and of Corollary 2.3 in the following ones. First, let us see that there exists a differentiable Jordan basis of the family of restrictions $A(t)| : Y_N(t) \longrightarrow Y_N(t)$. Let $\bar{w}_1(t), \ldots, \bar{w}_{n+m}(t)$ be a differentiable basis of $\mathbb{C}^{n+m}$ such that: $[\bar{w}_{s+1}(t), \ldots, \bar{w}_n(t)] = Y_N(t)$, $[\bar{w}_1(t), \ldots, \bar{w}_n(t)] = \mathbb{C}^n$, for all $t \in M$ (cf. Corollary 2.3). Because of $A(t)(Y_N(t)) \subset Y_N(t)$ and $B(t)(Y_N(t)) = 0$ for all $t \in M$, if we apply this change of basis to $\binom{A(t)}{B(t)}$ we obtain a matrix of the form

$$\begin{pmatrix} \overline{A}_{11}(t) & 0 \\ \overline{A}_{21}(t) & \overline{A}_{22}(t) \\ \overline{B}_1(t) & 0 \end{pmatrix}$$

where, for each $t \in M$, $\overline{A}_{22}(t)$, is the matrix of the restriction $A(t)|$ in the new basis $\bar{w}_{s+1}(t), \ldots, \bar{w}_n(t)$. According to hypothesis (ii′), $\overline{A}(t)$ has constant Jordan type, so that Theorem 3.2 can be applied. Hence, there exists a differentiable Jordan basis $w_{s+1}(t), \ldots, w_n(t)$ of $A(t)|$, and the first step is finished.

Let us denote $W(t)$ the differentiable basis $w_{s+1}(t), \ldots, w_n(t)$ of $Y_N(t)$ obtained above. Let us extend this basis successively. Corollary 2.3 ensures the existence of a differentiable basis $V_{N-1}(t)$ such that

$$Y_{N-1}(t) = [W(t)] \oplus [V_{N-1}(t)].$$

We have remarked in section 4.2 that map $A(t)$ is injective on $[V_{N-1}(t)]$, and that its image forms direct sum with $Y_{N-1}(t)$ in $Y_{N-2}(t)$. Then, again by virtue of Corollary 2.3, there exists a differentiable basis $V_{N-2}(t)$ such that

$$Y_{N-2}(t) = Y_{N-1}(t) \oplus [A(t)(V_{N-1}(t))] \oplus [V_{N-2}(t)].$$

Following this way, the next step gives

$$Y_{N-3}(t) = Y_{N-2}(t) \oplus [A^2(t)(V_{N-1}(t))] \oplus [A(t)(V_{N-2}(t))] \oplus [V_{N-3}(t)].$$

and so on. □

4.5. Additional remarks.

(1) In fact, Proposition 3.1 can be also applied to $\overline{A}(t)$ in the previous proof. Thus, in the conditions of Theorem 4.1

- there exist $\lambda_1, \ldots, \lambda_q \in C^p(M, \mathbb{C})$ such that $\lambda_1(t), \ldots, \lambda_q(t)$ are the distinct eigenvalues of $(A(t) \quad B(t))$, for every $t \in M$;
- the respective algebraic multiplicities $m_1, \ldots, m_q$ of these eigenvalues are constant.

(2) One has a similar result to the above theorem if we consider families of pairs of matrices $(A \quad B)$ with real coefficients. The only difference is that the matrix $J(t)$ in its Brunovsky form (cf. 4.1) is a real Jordan matrix instead of a complex one. Obviously, the corresponding matrices $S(t)$, $T(t)$, and $C(t)$ are also real.

5. Global pole assignment of class C^p. We assume that M is a contractible manifold, and $(A \;\; B) \in C^p(M, M_{n\times(n+m)}(\mathbb{R}))$, a differentiable family of pairs of matrices having constant Brunovsky type. Then, $s = k_1 + \cdots + k_r$ is constant. Also (cf. 4.5), there exist $\lambda_1, \ldots, \lambda_q \in C^p(M, \mathbb{C})$ such that $\lambda_1(t), \ldots, \lambda_q(t)$ are the distinct eigenvalues of $(A(t) \;\; B(t))$, having constant multiplicities $m_1, \ldots, m_q$.

We say that a set of maps $\mu_i \in C^p(M, \mathbb{C})$, $1 \le i \le s$, is *closed under conjugation* if for each i there is j such that $\mu_i(t) = \overline{\mu_j(t)}$ for every $t \in M$.

THEOREM 5.1. *Let M be a contractible manifold, $(A \;\; B) \in C^p(M, M_{n\times(n+m)}(\mathbb{R}))$ a differentiable family of pairs of matrices having constant Brunovsky type, $\lambda_1, \ldots, \lambda_q \in C^p(M, \mathbb{C})$ giving the distinct eigenvalues of $(A \quad B)$, and $m_1, \ldots, m_q$ their respective algebraic multiplicities. If $\mu_i \in C^p(M, \mathbb{C})$, $1 \le i \le s$, is a set of maps closed under conjugation, then there exists a family of matrices $K \in C^p(M, M_{m\times n}(\mathbb{R}))$ such that the eigenvalues of $A(t)+B(t)K(t)$ are $\mu_1(t), \ldots, \mu_s(t)$, $\lambda_1(t), \ldots, \lambda_q(t)$, the latter having multiplicities $m_1, \ldots, m_q$.*

Proof. As we have said above, it is a simple adaptation, by means of Theorem 4.1, of the proof in [9] relative to a constant pair. We enclose it for the convenience of the reader.

From section 4 we know that there exists $S \in C^p(M, Gl(n))$, $T \in C^p(M, Gl(m))$, $C \in C^p(M, M_{m\times n}(\mathbb{R}))$ such that

$$S(t)^{-1}(A(t) \quad B(t)) \begin{pmatrix} S(t) & 0 \\ C(t) & T(t) \end{pmatrix}$$

is a Brunovsky matrix for all $t \in M$. We denote by $(\bar{A} \;\; \bar{B})$ this Brunovsky family of matrices; that is to say,

$$\begin{aligned} \bar{A}(t) &= S(t)^{-1}(A(t)S(t) + B(t)C(t)), \\ \bar{B}(t) &= S(t)^{-1}B(t)T(t) \end{aligned}$$

We shall find a family $\bar{K} \in C^p(M, M_{m\times n}(\mathbb{R}))$ such that $\bar{A}(t) + \bar{B}(t)\bar{K}(t)$ has the desired eigenvalues $\mu_1(t), \ldots, \mu_s(t)$, $\lambda_1(t), \ldots, \lambda_q(t)$. Then, if we take $K(t) = C(t)S(t)^{-1} + T(t)\bar{K}(t)S(t)^{-1}$, the family $A(t) + B(t)K(t) = S(t)(\bar{A}(t) + \bar{B}(t)\bar{K}(t))S(t)^{-1}$ has the same eigenvalues. Let us construct $\bar{K}(t)$. One has

$$(\bar{A}(t) \quad \bar{B}(t)) = \begin{pmatrix} N_1 & & & & E_1 & & & \\ & \ddots & & & & \ddots & & \\ & & N_r & & & & E_r & \\ & & & J(t) & & & & 0 \end{pmatrix}$$

(see 4.1). Let $\ell_j = k_1 + \cdots + k_j$, $1 \le j \le r$, and $c_q^j \in C^p(M, \mathbb{C})$ defined by

$$(\xi - \mu_{\ell_{j-1}+1}(t)) \dots (\xi - \mu_{\ell_j}(t)) = \xi^{k_j} + \sum_{q=0}^{k_j-1} c_q^j(t)\xi^q,$$

where ξ is a indeterminate. Then,

$$(\xi - \mu_1(t)) \dots (\xi - \mu_s(t)) = \prod_{j=1}^{r} \left(\xi^{k_j} + \sum_{q=0}^{k_j-1} c_q^j(t)\xi^q \right),$$

and since $\{\mu_1, \dots, \mu_s\}$ is closed under conjugation we see that c_q^j are, in fact, real maps.

Let $\bar{K}(t)$ the $m \times n$ matrix defined by $\bar{K}(t) = (K_1(t)\ K_2(t)\ \dots\ K_r(t)\ 0)$ where

$$K_j(t) = \begin{pmatrix} & 0 & \\ -c_0^j(t) & \dots & -c_{k_j-1}^j(t) \\ & 0 & \end{pmatrix} (m\text{th row})$$

for $1 \le j \le r$. Then it is easy to see that the $n \times n$ matrix $\bar{A}(t) + \bar{B}(t)\bar{K}(t)$ is of the form

$$\begin{pmatrix} C_1(t) & & & & \\ & \ddots & & & \\ & & C_r(t) & & \\ & & & J(t) \end{pmatrix},$$

where

$$C_j(t) = \begin{pmatrix} & & \overset{(k_j)}{} & \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 \\ -c_0^j(t) & \dots & \dots & -c_{k_j-1}^j(t) \end{pmatrix}$$

Clearly, this matrix has the desired eigenvalues. □

Acknowledgments. We are grateful to the referees for their careful revision and their valuable suggestions.

REFERENCES

[1] V.I. Arnold, *On matrices depending on parameters*, Uspeckhi Mat. Nauk, 26 (1971), pp. 101–114.

[2] C.I. Byrnes, *On the control of certain deterministic, infinite-dimensional systems by algebra-geometric techniques*, Amer. J. Math., 100 (1978), pp. 1333–1381.

[3] L.D. Drager, R.L. Foote, and C.F. Martin, *Controllability of Linear Systems, Differential Geometry of Curves in Grassmannians, and Riccati Equations*, Tech. report, Texas Technical University, Dept. of Mathematics, Lubbock, TX, 1986.

[4] J.C. Evard and J.M. Gracia, *On similarities of class C^p and applications to matrix differential equations*, Linear Algebra Appl., 137 (1990), pp. 363–386.

[5] J. Ferrer, I. García, and F.Puerta, *Differentiable families of subspaces*, Linear Algebra Appl., 199 (1994), pp. 229–252.

[6] J. Ferrer and I. García, F.Puerta, *Brunovsky local form of a holomorphic family of pairs of matrices*, Linear Algebra Appl., 253 (1997), pp. 175–198.
[7] J. Ferrer and F. Puerta, *Similarity of non-everywhere defined linear maps*, Linear Algebra Appl., 168 (1992), pp. 27–55.
[8] R.M. Guralnick, *Similarity of matrices over commutative rings*, Linear Algebra Appl., 157 (1991), pp. 55–68.
[9] I. Gohberg, P. Lancaster, and L. Rodman, *Invariant Subspaces of Matrices with Applications*, John Wiley, New York, 1986.
[10] I. Gohberg and J. Leiterer, *Uber algebren stetiger operatorfunctionen*, Studia Math., LVII (1976), pp. 1–26.
[11] D. Husemoller, *Fibre Bundles*, Springer-Verlag, Berlin, 1975.
[12] M.L.J. Hautus and E.D. Sontag, *New results on pole-shifting for parametrized families of systems*, J. Pure Appl. Algebra, 40 (1986), pp. 229–244.
[13] M.Hazewinkel, *(Fine) moduli (spaces) for linear systems: What are they and what are they good for*, in Geometric Methods for the Theory of Linear Systems, C.I. Byrnes and C.F. Martin, eds., D. Reidel, Dordrecht, 1980.
[14] M. Hazewinkel and A.M.Perdon, *On families of systems: Pointwise-local-global isomorphism problems*, Internat. J. Control, 33 (1981), pp. 713–726.
[15] E.D. Sontag, *Linear Systems over Commutative rings: A survey*, Ricerche di Automatica, 7 (1976), pp. 1–34.
[16] E.D. Sontag, *On split realizations of reponse maps over rings*, Inform. and Control, 37 (1978), pp. 23–33.
[17] E.D. Sontag, *An introduction to the stabilization problem for parametrized families of linear systems*, Contemp. Math., 47 (1985), pp. 369–400.
[18] A. Tannenbaum, *Invariance and System Theory: Algebraic and Geometric Aspects,* Springer-Verlag, New York, 1980.
[19] A. Tannenbaum and P.P. Khargonekar, *On weak pole placement of linear systems depending on parameters*, in Mathematical Theory of Networks and Systems, P.A. Fuhrmann, ed., Springer-Verlag, Berlin, 1984, pp. 829-840.
[20] Y. Wang and E.D. Sontag, *Pole shifting for families of linear systems depending on at most three parameters*, Linear Algebra Appl., 138 (1990), pp. 3–38.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 933–955, October 1998

008

STABILITY AND CONVERGENCE OF PRINCIPAL COMPONENT LEARNING ALGORITHMS*

WEI-YONG YAN†

Abstract. This paper is concerned with the differential equation approximating the subspace learning algorithm for extracting principal components. Two issues are fully resolved. First, all the stable equilibria are found. Second, the global convergence rate is explicitly obtained. The whole treatment is without the nonsingularity assumption on the covariance matrix.

Key words. stability, convergence, matrix analysis, neural networks

AMS subject classifications. 34D05, 65L20, 15A18, 15A03

PII. S0895479896310329

1. Introduction. As one of the most important principal component analysis techniques, the subspace algorithm has been used for performing several different learning tasks in the context of linear neural networks, see, e.g., [1, 2, 3, 4, 5]. The implementation of this algorithm is fairly straightforward. From a system point of view, the algorithm is simply a nonlinear system, where the input is a random signal within some class and the output is a connection or synaptic weight matrix. Precisely, the algorithm is described by the recursive equation

$$W_{k+1} = W_k + \gamma(x_k - W_k y_k)y_k^T,$$
$$y_k = W_k^T x_k,$$

which is associated with and can be approximated in an average sense by the ordinary differential equation (ODE)

$$\dot{W}(t) = [I - W(t)W^T(t)]CW(t),$$

obtained by Oja [3], where C is the covariance matrix of the input signal. Most remarkably, such an algorithm, which is composed of the Hebbian rule and additional feedback, is capable of extracting the main features of the input signal class.

The mathematical analysis of the subspace algorithm, which has proved to be difficult due to the nonlinear complexity, can be traced back to Oja's early work [6] in the single neuron case. Later, in [7], Oja and Karhunen established a more precise connection between the one-unit algorithm and the associated ODE. The validity of the orthonormalization capability in the multineuron case was found by Oja [3] from a simulation study. Theoretically, there has been considerable effort made in order to gain insight into the subspace algorithm. For example, the approximation of the subspace algorithm by the ODE was rigorously developed in detail by Hornik and Kuan [8]. A local analysis of the associated ODE was given by Williams [1] and by Krogh and Hertz [9]. A close form time-domain solution can be found in [10] and an error function for the ODE in [11]. Recently, the global convergence analysis has been completed by Yan, Helmke, and Moore [10], where all Oja's conjectures are

*Received by the editors October 7, 1996; accepted for publication (in revised form) by M. Chu September 25, 1997; published electronically July 6, 1998.
http://www.siam.org/journals/simax/19-4/31032.html
†School of Electrical and Computer Engineering, Curtin University of Technology, GPO Box U1987, Pert, Western Australia 6845, Australia (wyy@cs.curtin.edu.au).

rigorously proved when the input covariance matrix is nonsingular. It is interesting that the above ODE has the same form as the gradient flow of the generalized Rayleith quotient on the Stiefel manifold of real orthogonal matrices. The global convergence of such a flow with an initial condition restricted to being in the Stiefel manifold has been studied in [12] using differential manifold techniques.

In this paper, we tackle some remaining theoretic issues surrounding the subspace algorithm. Perhaps the stability issue is most important among them. Though it is known that the associated ODE cannot have any asymptotically stable equilibrium since no equilibrium is isolated, it is unclear whether there are any stable equilibria and how to find them if there are any. The second important issue is related to the quantitative performance of the subspace algorithm. Obviously, the performance is best measured by the convergence rate, which has not been available so far. The exponential convergence rate to be derived in the paper in terms of the eigenvalues of the covariance matrix will also enable us to complete the global convergence analysis without requiring the assumption that the input covariance matrix is nonsingular.

The layout of the paper is quite straightforward. Following some technical preparations in the next section, the global convergence analysis will first be carried out in section 3. Then we shall discuss the extraction of dominant eigenspaces in section 4 and proceed to resolve the stability issue in section 5. Finally, conclusions are drawn in section 6.

This introduction is ended with a list of some mathematical symbols and matrix notation to be used.

- $O(e^{\alpha t})$ stands for any function $f(t)$ with the property that $e^{-\alpha t}f(t)$ is bounded for $t > 0$.
- $A \geq B$ means that $A - B$ is a nonnegative definite symmetric matrix.
- $A^{1/2}$ — the nonnegative definite square root of $A \geq 0$.
- $\mathbb{R}^{n\times m}$ — the set of all $n \times m$ real matrices.
- $\exp(A)$ — the matrix exponential of A.
- $\|A\|$ — the spectral norm, i.e., the maximum singular value of A.
- A^T — the transpose of A.
- $A^\dagger$ — the pseudoinverse of A.
- I — an identity matrix of appropriate dimensions.
- I_n — the $n \times n$ identity matrix.
- range A — the subspace spanned by the columns of A.
- ker A — the null subspace of A.

2. Preliminary results. In this section, we present some technical lemmas for later use. The first lemma is given without proof.

LEMMA 2.1. *Given $A, B \in \mathbb{R}^{n\times n}$ with $A, B \geq 0$, and $X \in \mathbb{R}^{m\times n}$.*

(i) *If*

$$XA^iB = 0, \quad \forall\, i = 0,\ 1, \ldots, n,$$

then

$$X(A + B)^{1/2} = XA^{1/2.}$$

(ii) *There holds*

$$\left\|A^{1/2} - B^{1/2}\right\| \leq \|A - B\|^{1/2}.$$

LEMMA 2.2. *Let $\Delta \in \mathbb{R}^{n\times n}$ be given with $\Delta \geq 0$. Then for any $X \in \mathbb{R}^{n\times m}$ there holds*

$$\left\|(I + X^T\Delta^{-1}X)^{-1} - [I - X^T\left(XX^T\right)^\dagger X]\right\| \leq \left\|X^\dagger\right\|^2 \|\Delta\|.$$

Proof. See Appendix A. □

The next two lemmas are crucial not only to our derivation of the exponential convergence rate but also to the stability analysis.

LEMMA 2.3. *Let $X_i \in \mathbb{R}^{m_i\times k}$ and $\Delta_i \in \mathbb{R}^{m_i\times m_i}$ with $\Delta_i \geq 0$ for $i = 1, 2$. Assume that there are two positive constants α_1 and α_2 such that*

$$\Delta_1 \geq \alpha_1 I \geq \alpha_2 I \geq \Delta_2.$$

Then there holds

$$\left\|\Delta_1 X_1\left[\left(I + X_1^T\Delta_1^2X_1 + X_2^T\Delta_2^2X_2\right)^{-1/2} - \left(I + X_1^T\Delta_1^2X_1\right)^{-1/2}\right]\right\| \leq 2(\alpha_2/\alpha_1)\|X_2\|\left\|X_1^\dagger\right\|.$$

Proof. See Appendix A. □

LEMMA 2.4. *Let $C_1 \geq 0$ be a constant matrix with the minimum eigenvalue c_1 and let c_2 be a constant scalar. Assume that $X_2 \neq 0$. If $c_1 > c_2 > 0$, then for $t > 0$ there holds*

$$\begin{aligned}\Big\|\exp(c_2t)X_2\left\{I + X_1^T[\exp(2C_1t) - I]X_1 + [\exp(2c_2t) - 1]X_2^TX_2\right\}^{-1/2} \\ - X_2\left[S - SX_1^T\left(X_1SX_1^T\right)^\dagger X_1S\right]^{1/2}\Big\| \\ \leq \|X_2\|\left(1 + \left\|X^\dagger\right\|^2 + \beta\|X\|\left\|X^\dagger\right\|\left\|X_1^\dagger\right\|\right)e^{-\alpha t},\end{aligned}$$

where

$$S = \left(X_1^TX_1 + X_2^TX_2\right)^\dagger,$$
$$\alpha = \min(c_2,\ c_1 - c_2), \quad \beta = \sqrt{\frac{c_1}{c_1 - c_2}}.$$

Proof. See Appendix A. □

3. Global convergence. The global convergence of the subspace algorithm was first discovered numerically by Oja [3] and was theoretically proved by Yan, Helmke, and Moore [10] in the case where the input covariance matrix is nonsingular. The objective of this section is twofold. First, it will be proved that the global convergence remains true for the general case. Second, the convergence rate will be derived.

Recall that the subspace algorithm for learning principal subspaces is approximated by the ODE equation

$$\dot{W} = (I - WW^T)CW, \quad W(0) = W_0, \tag{3.1}$$

where $C \in \mathbb{R}^{n\times n}$ is the input covariance matrix and $W_0 \in \mathbb{R}^{n\times k}$ is the initial weight matrix with k being the number of the neurons used in learning. It will prove useful to have a singular value decomposition (SVD) of C

$$C = U\,\mathrm{diag}\,c_1I_{n_1},\ c_2I_{n_2}, \dots,\ c_pI_{n_p},\ 0I_{n_{p+1}}U^T \tag{3.2}$$

with

$$U = \begin{bmatrix} U_1 & U_2 & \cdots & U_p & U_{p+1} \end{bmatrix}, \tag{3.3}$$

where U is orthogonal, $U_i \in \mathbb{R}^{n \times n_i}$, $c_1 > c_2 > \cdots c_p > 0$.

Now define the matrix

$$\Theta \triangleq U \begin{bmatrix} U_1^T W_0 V_1 \\ U_2^T W_0 V_2 \\ \vdots \\ U_{p+1}^T W_0 V_{p+1} \end{bmatrix} \tag{3.4}$$

with

$$V_i = \left\{ \left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger - \left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \mathcal{U}_{i-1}^T \left[\mathcal{U}_{i-1} \left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \mathcal{U}_{i-1}^T \right]^\dagger \mathcal{U}_{i-1} \left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \right\}^{1/2} \tag{3.5}$$

and

$$\mathcal{U}_i^T = \begin{cases} 0 & i = 0, \\ W_0^T \begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix} & 1 \le i \le p, \\ I & i = p+1. \end{cases} \tag{3.6}$$

As will be seen, the columns of the solution $W(t)$ to (3.1) tend to span the same subspace as the columns of the matrix Θ as t goes to infinity. Some interesting properties regarding the above matrices are given below.

LEMMA 3.1. *Let V_i be defined in terms of W_0 via* (3.5)–(3.6) *for $i = 1, 2, \ldots, p+1$. Then the following relations hold:*

(i)

$$\ker V_i = \operatorname{range} \mathcal{U}_{i-1}^T \oplus \ker \mathcal{U}_i,$$

(ii)

$$(U_i^T W_0 V_i)^T (U_i^T W_0 V_i) = V_i V_i^\dagger, \quad i < p+1,$$

(iii)

$$V_i V_j = 0, \quad i \neq j,$$

(iv)

$$U_i^T W_0 V_i^2 W_0^T U_i = I \quad \Longleftrightarrow \quad \operatorname{rank} \mathcal{U}_i = n_i + \operatorname{rank} \mathcal{U}_{i-1}.$$

Proof. See Appendix A. ∎

A key to our subsequent development is the following representation of the solution $W(t)$.

LEMMA 3.2 (see [10]). *The solution $W(t)$ to* (3.1) *obeys*

$$W(t)W^T(t) = \exp(Ct) W_0 [I_k - W_0^T W_0 + W_0^T \exp(2Ct) W_0]^{-1} W_0^T \exp(Ct)$$

and

$$\|W(t)\| \leq \max(1,\ \|W_0\|), \quad \forall\, t \geq 0.$$

We are now in a position to present one of our main results concerning the exponential convergence rate for (3.1).

THEOREM 3.1. *Let $W(t)$ be the solution to the ODE* (3.1) *and define Θ as in* (3.4). *Then there exists some constant orthogonal matrix $\Pi \in \mathbb{R}^{k\times k}$ such that*

$$\|W(t) - \Theta\Pi\| = O\left(e^{-\mu t}\right), \tag{3.7}$$

where $\mu > 0$ is defined by

$$\mu \triangleq \min(c_p,\ c_1 - c_2,\ c_2 - c_3,\ \ldots,\ c_{p-1} - c_p).$$

Moreover, $X = \Theta\Theta^T$ satisfies the following relation:

$$CX = XC = XCX. \tag{3.8}$$

Proof. Equation (3.8) is obvious upon noting that

$$\begin{aligned} C\Theta\Theta^T &= \Theta\Theta^T C = \Theta\Theta^T C\Theta\Theta^T \\ &= U \begin{bmatrix} c_1(U_1^T W_0)V_1^2(U_1^T W_0)^T & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & c_p\left(U_p^T W_0\right) V_p^2 \left(U_p^T W_0\right)^T & 0 \\ 0 & \cdots & 0 & 0 \end{bmatrix} U^T, \end{aligned}$$

which follows from (ii) and (iii) of Lemma 3.1. So, it suffices to prove (3.7). With

$$F(t) \triangleq \exp(Ct)W_0[I_k - W_0^T W_0 + W_0^T \exp(2Ct)W_0]^{-1/2},$$

it is seen from Lemma 3.2 that $F(t)$ is bounded with respect to $t \geq 0$ and that $W(t)W^T(t) = F(t)F^T(t)$, which implies that there exists an orthogonal matrix $Z(t)$ such that

$$W(t) = F(t)Z(t), \quad \forall\, t \geq 0. \tag{3.9}$$

In addition, note that $F(t)$ can be rewritten as

$$F(t) = U \begin{bmatrix} e^{c_1 t}U_1^T W_0 \\ \vdots \\ e^{c_p t}U_p^T W_0 \\ U_{p+1}^T W_0 \end{bmatrix} \left[I + \sum_{i=1}^{p} \left(e^{2c_i t} - 1\right)\left(U_i^T W_0\right)^T U_i^T W_0 \right]^{-1/2}. \tag{3.10}$$

Now, it is inferred by Lemma 2.3 that for $t > 0$,

$$\begin{aligned}
&\left\| e^{c_j t} U_j^T W_0 \left\{ \left[I + \sum_{i=1}^{p} \left(e^{2c_i t} - 1 \right) \left(U_i^T W_0 \right)^T U_i^T W_0 \right]^{-1/2} \right.\right. \\
&\qquad \left.\left. - \left[I + \sum_{i=1}^{j} \left(e^{2c_i t} - 1 \right) \left(U_i^T W_0 \right)^T U_i^T W_0 \right]^{-1/2} \right\} \right\| \\
&\le 2 \left\| \mathcal{U}_j^\dagger \right\| \left\| W_0^T \left[U_{j+1} \quad \cdots \quad U_p \right] \right\| \frac{e^{c_{j+1} t} - 1}{e^{c_j t} - 1} \\
&\quad \left\| U_j^T W_0 \left\{ \left[I + \sum_{i=1}^{p} \left(e^{2c_i t} - 1 \right) \left(U_i^T W_0 \right)^T U_i^T W_0 \right]^{-1/2} \right.\right. \\
&\qquad \left.\left. - \left[I + \sum_{i=1}^{j} \left(e^{2c_i t} - 1 \right) \left(U_i^T W_0 \right)^T U_i^T W_0 \right]^{-1/2} \right\} \right\| \\
&\le 2 \left\| W_0 \right\| \left\| \mathcal{U}_j^\dagger \right\| e^{-(c_j - c_{j+1})t} + 2 e^{-c_j t}
\end{aligned} \tag{3.11}$$

and by Lemma 2.4 that

$$\begin{aligned}
&\left\| e^{c_j t} U_j^T W_0 \left[I + \sum_{i=1}^{j} \left(e^{2c_i t} - 1 \right) \left(U_i^T W_0 \right)^T U_i^T W_0 \right]^{-1/2} - U_j^T W_0 V_j \right\| \\
&\le \left\| U_j^T W_0 \right\| \left(1 + \left\| \mathcal{U}_j^\dagger \right\|^2 + \beta \left\| \mathcal{U}_j \right\| \left\| \mathcal{U}_j^\dagger \right\| \left\| \mathcal{U}_{j-1}^\dagger \right\| \right) e^{-\min(c_{j-1} - c_j,\, c_j)t}
\end{aligned} \tag{3.12}$$

with $\beta = \sqrt{\frac{c_{j-1}}{c_{j-1} - c_j}}$. The combination of (3.11) and (3.12) results in

$$\begin{aligned}
&\left\| e^{c_j t} U_j^T W_0 \left[I + \sum_{i=1}^{p} \left(e^{2c_i t} - 1 \right) \left(U_i^T W_0 \right)^T U_i^T W_0 \right]^{-1/2} - U_j^T W_0 V_j \right\| \\
&= O\left(e^{-\min(c_{j-1} - c_j,\, c_j - c_{j+1})t} \right), \quad j = 1, \ldots, p.
\end{aligned}$$

Meanwhile, from (ii) of Lemma 2.1 and Lemma 2.2 one has

$$\begin{aligned}
&\left\| U_{p+1}^T W_0 \left[I + \sum_{i=1}^{p} \left(e^{2c_i t} - 1 \right) \left(U_i^T W_0 \right)^T U_i^T W_0 \right]^{-1/2} - U_{p+1}^T W_0 V_{p+1} \right\| \\
&\le \left\| W_0 \right\| \left\| \left[I + \sum_{i=1}^{p} \left(e^{2c_i t} - 1 \right) \left(U_i^T W_0 \right)^T U_i^T W_0 \right]^{-1} - \left[I - \mathcal{U}_p^T \left(\mathcal{U}_p \mathcal{U}_p^T \right)^\dagger \mathcal{U}_p \right] \right\|^{1/2} \\
&\le \left\| W_0 \right\| \left[\left\| \mathcal{U}_p^\dagger \right\|^2 \left(e^{2c_p t} - 1 \right)^{-1} \right]^{1/2} \\
&= O\left(e^{-c_p t} \right).
\end{aligned} \tag{3.13}$$

It is thus seen from (3.10) that

$$\|F(t) - \Theta\| = O\left(e^{-\mu t}\right). \tag{3.14}$$

Due to (3.8), one has

$$(I - \Theta\Theta^T)C\Theta = 0.$$

As a consequence, it follows from (3.9) and (3.14) that

$$\begin{aligned}\left\|\dot{W}(t)\right\| =& \left\|[I - W(t)W^T(t)]CW(t) - (I - \Theta\Theta^T)C\Theta Z(t)\right\| \\ \leq& \left\|[F(t)F^T(t) - \Theta\Theta^T]CW(t)\right\| + \left\|(I - \Theta\Theta^T)C[F(t) - \Theta]\right\| \\ =& O\left(e^{-\mu t}\right).\end{aligned}$$

Namely, there is some constant $K > 0$ such that

$$\left\|\dot{W}(t)\right\| \leq Ke^{-\mu t}.$$

With this, it is true that for $t_2 > t_1 \geq 0$,

$$\|W(t_2) - W(t_1)\| \leq \int_{t_1}^{t_2} \left\|\dot{W}(t)\right\| dt \leq \int_{t_1}^{t_2} Ke^{-\mu t} dt = \frac{K}{\mu}\left(e^{-\mu t_1} - e^{-\mu t_2}\right),$$

which clearly shows the existence of $\lim_{t\to\infty} W(t)$ by the Cauchy criterion. Denoting this limit by W_∞ and letting t_2 go to infinity in the above gives rise to

$$\|W(t_1) - W_\infty\| \leq \frac{K}{\mu} e^{-\mu t_1}.$$

W_∞ must be of the form $\Theta\Pi$ for some orthogonal matrix Π due to

$$W_\infty W_\infty^T = \lim_{t\to\infty} F(t)F^T(t) = \Theta\Theta^T.$$

The theorem is thus proved. □

From now on, we call the limit of the solution $W(t)$ as $t \to \infty$ the limiting solution to (3.1). The following result directly follows from the foregoing theorem.

COROLLARY 3.1. *Consider the ODE* (3.1) *with an arbitrarily given initial weight matrix* W_0. *Then there exists some constant orthogonal matrix* $\Pi \in \mathbb{R}^{k\times k}$ *such that*

$$\lim_{t\to\infty} W^T(t)W(t) = \Pi^T\Theta^T\Theta\Pi,$$

where Θ *is defined as in* (3.4).

Remark 3.1. Note that the matrix Θ has a simpler form for a generic initial point W_0 when the covariance matrix C has distinct eigenvalues. In fact, all the U_i are column vectors in this case. If W_0 satisfies the rank condition

$$\operatorname{rank} W_0^T \begin{bmatrix} U_1 & U_2 & \cdots & U_k \end{bmatrix} = k; \tag{3.15}$$

then there holds

$$\operatorname{range} \mathcal{U}_i^T = \mathbb{R}^k, \quad \forall\, i \geq k,$$

which implies by (i) of Lemma 3.1 that

$$U_i^T W_0 V_i = 0, \quad \forall\, i > k.$$

This leads to

$$\Theta = \begin{bmatrix} U_1 & U_2 & \cdots & U_k \end{bmatrix} \begin{bmatrix} U_1^T W_0 V_1 \\ U_2^T W_0 V_2 \\ \vdots \\ U_k^T W_0 V_k. \end{bmatrix}$$

Moreover, from (iii) and (iv) of Lemma 3.1 it can be further seen that $\Theta^T\Theta$ is an identity matrix. As a consequence, it is concluded that $W^T(t)W(t)$ converges to the identity matrix for any initial point W_0 satisfying (3.15) in the case where C has distinct eigenvalues.

4. Extraction of principal subspaces. Having established the global convergence for the subspace equation, we turn to discuss generic properties of the limiting solution associated with a given initial condition. In particular, we will examine when the subspace algorithm can extract a dominant eigenspace for a generic initial weight matrix. Throughout, W_∞ will stand for the limiting solution to the ODE (3.1).

DEFINITION 4.1. *Let C be a covariance matrix with the SVD* (3.2)–(3.3). *The subspace spanned by the columns of the matrix $\begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix}$ is called a dominant eigenspace of C, where i is any integer with $1 \le i \le p+1$.*

DEFINITION 4.2. *Let a property* **P** *depend on a variable x in a real n-dimensional Euclidean space. A subset of the space is called a proper variety if there exists a finite system of polynomial equations in n-indeterminates, in which at least one polynomial is nonzero, such that every element of the subset is a zero of the system.* **P** *is said to be true for almost all x in the space if all the x for which* **P** *does not hold belong to a proper variety of the space.*

Our first result in this section states that both a given initial weight matrix and the resulting limiting solution W_∞ obey a series of rank conditions expressed in terms of the eigenvectors of the covariance matrix. This result will be used to characterize the set of initial weight matrices which lead to stable equilibria in the next section.

THEOREM 4.1. *Consider the ODE* (3.1) *with the initial condition $W(0) = W_0$. Then there holds*

$$\operatorname{rank} W_\infty^T \begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix} = \operatorname{rank} W_0^T \begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix} \tag{4.1}$$

for $i = 1, \ldots, p$.

Proof. First, it can be seen from (i) of Lemma 3.1 that

$$\dim \ker V_i = \dim \operatorname{range} \mathcal{U}_{i-1}^T + \dim \ker \mathcal{U}_i,$$

which implies that

$$\operatorname{rank} V_i = \operatorname{rank} \mathcal{U}_i - \operatorname{rank} \mathcal{U}_{i-1}, \tag{4.2}$$

where V_i and $\mathcal{U}_i$ are defined as before. Next, by appealing to Theorem 3.1 and (ii) of Lemma 3.1 one obtains

$$\begin{aligned} & \begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix}^T W_\infty W_\infty^T \begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix} \\ = & \begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix}^T \Theta\Theta^T \begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix} \\ = & \text{ block diag} \left\{ V_1 V_1^\dagger, \cdots, V_i V_i^\dagger \right\}, \end{aligned}$$

which implies that

$$\operatorname{rank} W_\infty^T \begin{bmatrix} U_1 & U_2 & \cdots & U_i \end{bmatrix} = \sum_{j=1}^{i} \operatorname{rank} V_j.$$

Combining this with (4.2) immediately yields (4.1). ∎

The following result indicates that the number of neurons employed in a network may affect the capability of extracting a principal subspace of an input signal. To ensure this capability, the number of neurons k must be such that the kth largest eigenvalue of the covariance matrix is strictly greater than the $(k+1)$th largest one.

THEOREM 4.2. *Consider the subspace equation* (3.1) *for the case of k neurons. Let the input covariance matrix C have the eigenvalues*

$$l_1 \geq l_2 \geq \cdots \geq l_n$$

and W_∞ be the limiting solution to (3.1).

(1) *If $l_k > l_{k+1}$, the columns of W_∞ span a k-dimensional dominant eigenspace for almost all initial weight matrices $W_0 \in \mathbb{R}^{n\times k}$.*

(2) *If $l_k = l_{k+1}$, the range of W_∞ is a direct sum of a dominant eigenspace and one nonempty proper subspace of the eigenspace corresponding to the eigenvalue l_k for almost all initial weight matrices $W_0 \in \mathbb{R}^{n\times k}$.*

Proof. As before, we adopt the SVD of C as in (3.2), where $c_1 > c_2 > \cdots > c_p$. Then evidently, there holds $l_k = c_r$ for some integer r with

$$1 \leq r \leq p.$$

Accordingly, it is true that

$$\sum_{i=1}^{r-1} n_i < k \leq \sum_{i=1}^{r} n_i.$$

Now with the assumption that $l_k > l_{k+1}$, the number of columns of the matrix

$$\begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix} \tag{4.3}$$

must equal k. So the condition

$$\operatorname{rank} W_0^T \begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix} = k \tag{4.4}$$

holds for almost all $W_0 \in \mathbb{R}^{n\times k}$. Moreover, under this condition, it follows from (i) and (iv) of Lemma 3.1 that

$$U_i^T W_0 V_i^2 W_0^T U_i = \begin{cases} I, & i = 1, \ldots, r, \\ 0, & i = r+1, \ldots, p+1, \end{cases}$$

which implies that

$$W_\infty W_\infty^T = \begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix} \begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix}^T.$$

In particular, there holds

$$\operatorname{range} W_\infty = \operatorname{range} \begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix}.$$

Therefore, (1) is proved. Next, we assume that $l_k = l_{k+1}$, in which case the number of columns of the matrix in (4.3) is obviously greater than k, i.e.,

$$\operatorname{rank}\begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix} > k. \tag{4.5}$$

If there holds

$$\operatorname{rank} W_0^T \begin{bmatrix} U_1 & U_2 & \cdots & U_{r-1} \end{bmatrix} = \sum_{i=1}^{r-1} n_i, \tag{4.6}$$

$$\operatorname{rank} W_0^T \begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix} = k, \tag{4.7}$$

then it follows again from (i) and (iv) of Lemma 3.1 that

$$U_i^T W_0 V_i^2 W_0^T U_i = \begin{cases} I, & i = 1, \ldots, r-1, \\ 0, & i = r+1, \ldots, p+1, \end{cases}$$

leading to

$$\begin{aligned} &W_\infty W_\infty^T \\ &= \begin{bmatrix} U_1 & U_2 & \cdots & U_{r-1} & U_r U_r^T W_0 V_r \end{bmatrix} \begin{bmatrix} U_1 & U_2 & \cdots & U_{r-1} & U_r U_r^T W_0 V_r \end{bmatrix}^T. \end{aligned}$$

It is thus deduced that

$$\operatorname{range} W_\infty = \operatorname{range} \begin{bmatrix} U_1 & U_2 & \cdots & U_{r-1} \end{bmatrix} \oplus \operatorname{range} U_r U_r^T W_0 V_r.$$

Note that the range of $U_r U_r^T W_0 V_r$ cannot be equal to that of U_r due to (4.5) and $W_\infty \in \mathbb{R}^{n\times k}$. Moreover, it is not difficult to see from (i) of Lemma 3.1 that $U_r^T W_0 V_r$ cannot be zero unless

$$\operatorname{range} W_0^T U_r \subset \operatorname{range} W_0^T \begin{bmatrix} U_1 & U_2 & \cdots & U_{r-1} \end{bmatrix},$$

which is impossible because of (4.7). In summary, the range W_∞ is a direct sum of the range of $\begin{bmatrix} U_1 & U_2 & \cdots & U_{r-1} \end{bmatrix}$ and a *nonempty proper* subspace of the range of U_r under the two conditions (4.6) and (4.7). Since these two conditions hold for almost all $W_0 \in \mathbb{R}^{n\times k}$, (2) is concluded. □

5. Stability. Concerning the stability of the subspace equation, there are two known facts: one is that there exists no asymptotically stable equilibrium and the other is that any equilibrium whose range is perpendicular to a dominant eigenspace is unstable. The first fact is plain because no equilibrium is isolated while the second one is intuitively clear by recalling that the solution to the equation tends to span a dominant eigenspace for almost all initial points. In this section, it will be seen that all the stable equilibria can be found and parameterized in a simple and explicit way. In addition, we shall also identify the class of perturbations about an unstable equilibrium, which do not lead the solution to deviate radically from the equilibrium.

Recall that the ODE associated with the subspace algorithm is given by

$$\dot{W} = (I - WW^T)CW. \tag{5.1}$$

Without loss of generality, it will be assumed throughout this section that the covariance matrix $C \in \mathbb{R}^{n\times n}$ is nonzero and that the weight matrix W is $n \times k$ with $n \geq k$.

Let $\mathfrak{E}$ denote the set of all the equilibrium points of the associated ODE. In other words, $\mathfrak{E}$ is composed of all the solutions to the algebraic equation

$$(I - WW^T)CW = 0, \tag{5.2}$$

i.e.,

$$\mathfrak{E} = \{W \in \mathbb{R}^{n\times k};\ (I - WW^T)CW = 0\},$$

which is a closed set in $\mathbb{R}^{n\times k}$. A characterization of this set was given by Oja in [13] when C is positive definite. To get an explicit parametrization of $\mathfrak{E}$ in the general case, we decompose as before the input covariance matrix C as

$$C = U \operatorname{diag} c_1 I_{n_1},\ c_2 I_{n_2}, \dots, c_p I_{n_p},\ 0 I_{n_{p+1}} U^T \tag{5.3}$$

with

$$U = \begin{bmatrix} U_1 & U_2 & \cdots & U_p & U_{p+1} \end{bmatrix},$$

where U is orthogonal, $U_i \in \mathbb{R}^{n\times n_i}$, $c_1 > c_2 > \cdots c_p > 0$. Then by Theorem 3.1, $\mathfrak{E}$ can be parameterized in terms of an arbitrary W_0 and arbitrary orthogonal Π as follows:

$$\mathfrak{E} = \left\{ \sum_{i=1}^{p+1} U_i U_i^T W_0 V_i \Pi;\ \ W_0 \in \mathbb{R}^{n\times k} \quad \text{and} \quad \Pi \in \mathbb{R}^{k\times k},\ \Pi^T\Pi = I \right\},$$

where V_i is defined via (3.5) and (3.6).

For the purpose of identifying all the stable equilibria out of $\mathfrak{E}$, introduce the two complementing subsets of $\mathfrak{E}$:

$$\mathfrak{E}_s \triangleq \{W \in \mathfrak{E};\ (5.6) \text{ and } (5.7) \text{ both hold:}\}, \tag{5.4}$$

$$\mathfrak{E}_u \triangleq \mathfrak{E} - \mathfrak{E}_s \tag{5.5}$$

with

$$\operatorname{rank} \begin{bmatrix} U_1 & \cdots & U_{r-1} \end{bmatrix}^T W = \sum_{i=1}^{r-1} n_i, \tag{5.6}$$

$$\operatorname{rank} \begin{bmatrix} U_1 & \cdots & U_r \end{bmatrix}^T W = k, \tag{5.7}$$

where r is the unique index such that the kth largest eigenvalue of C equals c_r.

Remark 5.1. Quite obviously, r satisfies the inequality

$$\sum_{i=1}^{r-1} n_i < k \le \sum_{i=1}^{r} n_i, \tag{5.8}$$

where $\sum_{i=1}^{r-1} n_i$ will be understood to be zero if $r = 1$. Moreover, k is equal to $\sum_{i=1}^{r} n_i$ if and only if the kth largest eigenvalue of C is strictly larger than the $(k+1)$th largest one.

As will be seen shortly, the set $\mathfrak{E}_s$ actually contains all the stable equilibria of the subspace equation. In view of this, we shall first try to simplify the representation of $\mathfrak{E}_s$. As a direct consequence of Theorem 4.1, the following lemma is obtained which

gives a parametrization of the set $\mathfrak{E}_s$ by characterizing the set of initial points leading to equilibria in $\mathfrak{E}_s$.

LEMMA 5.1. *Consider the ODE* (5.1). *Then the set* $\mathfrak{E}_s$ *defined as in* (5.4) *is given by*

$$\mathfrak{E}_s = \left\{ \sum_{i=1}^{p+1} U_i U_i^T W_0 V_i \Pi;\ \ W_0 \in \mathbb{R}^{n\times k} \text{ satisfies (5.10)–(5.11), and } \Pi \in \mathbb{R}^{k\times k},\ \Pi^T\Pi = I \right\} \tag{5.9}$$

with

$$\operatorname{rank}\begin{bmatrix} U_1 & \cdots & U_{r-1} \end{bmatrix}^T W_0 = \sum_{i=1}^{r-1} n_i, \tag{5.10}$$

$$\operatorname{rank}\begin{bmatrix} U_1 & \cdots & U_r \end{bmatrix}^T W_0 = k. \tag{5.11}$$

Alternatively and perhaps more elegantly, the set $\mathfrak{E}_s$ can be explicitly parameterized in terms of two independent orthogonal matrices without reference to the initial point W_0.

THEOREM 5.1. *Let* r *be the unique index such that* c_r *is the* k*th largest eigenvalue of* C *and define the set* $\mathfrak{E}_s$ *as in* (5.4). *If* $r < p+1$, *then*

$$\mathfrak{E}_s = \Big\{ \begin{bmatrix} U_1 & \cdots & U_{r-1} & U_r\Lambda \end{bmatrix} \Pi; \quad \Lambda \in \mathbb{R}^{n_r \times (k - \sum_{i=1}^{r-1} n_i)},\ \Pi \in \mathbb{R}^{k\times k},\ \Lambda^T\Lambda = I,\ \Pi^T\Pi = I \Big\}; \tag{5.12}$$

otherwise,

$$\mathfrak{E}_s = \left\{ \begin{bmatrix} U_1 & \cdots & U_{r-1} & U_r\Lambda \end{bmatrix} \Pi; \quad \Lambda \in \mathbb{R}^{n_r \times (k - \sum_{i=1}^{r-1} n_i)},\ \Pi \in \mathbb{R}^{k\times k},\ \operatorname{rank}\Lambda = k - \sum_{i=1}^{r-1} n_i,\ \Pi^T\Pi = I \right\}. \tag{5.13}$$

Proof. First it can be directly verified that the algebraic equation (5.2) and the two rank conditions (5.6)–(5.7) are satisfied by any matrix of the form

$$\begin{bmatrix} U_1 & \cdots & U_{r-1} & U_r\Lambda \end{bmatrix} \Pi$$

with

$$\Lambda \in \mathbb{R}^{n_r \times (k - \sum_{i=1}^{r-1} n_i)},\ \Pi \in \mathbb{R}^{k\times k},\ \operatorname{rank}\Lambda = k - \sum_{i=1}^{r-1} n_i,\ \Pi^T\Pi = I.$$

Thus, it is true that $\mathfrak{E}_s$ includes as a subset the set on the right side of (5.12) or (5.13), depending on whether $r < p+1$.

To prove the converse inclusion, let $W_e \in \mathfrak{E}_s$. Namely, W_e satisfies (5.2), (5.6), and (5.7). It is apparent that if W_e is used as an initial point of the ODE (5.1), then

the limiting solution W_∞ as well as the solution $W(t)$ will equal W_e. By Theorem 3.1, there is some orthogonal matrix Π_1 such that

$$W_e = U \begin{bmatrix} U_1^T W_e V_1 \\ U_2^T W_e V_2 \\ \vdots \\ U_{p+1}^T W_e V_{p+1} \end{bmatrix} \Pi_1.$$

Since every row of $U_i^T W_e$ with $i > r$ is a linear combination of the rows of the matrix

$$\begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix}^T W_e,$$

which is of full column rank, it follows from (i) of Lemma 3.1 that $U_i^T W_e V_i = 0$ for $i > r$, leading to

$$W_e = \begin{bmatrix} U_1 & U_2 & \cdots & U_r \end{bmatrix} \begin{bmatrix} U_1^T W_e V_1 \\ U_2^T W_e V_2 \\ \vdots \\ U_r^T W_e V_r \end{bmatrix} \Pi_1. \tag{5.14}$$

Since $W = W_e$ satisfies (5.6), it follows from (iii) and (iv) of Lemma 3.1 that

$$\begin{bmatrix} U_1^T W_e V_1 \\ \vdots \\ U_{r-1}^T W_e V_{r-1} \\ U_r^T W_e V_r \end{bmatrix} \begin{bmatrix} U_1^T W_e V_1 \\ \vdots \\ U_{r-1}^T W_e V_{r-1} \\ U_r^T W_e V_r \end{bmatrix}^T = \begin{bmatrix} I & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I & 0 \\ 0 & \cdots & 0 & U_r^T W_e V_r^2 W_e^T U_r \end{bmatrix}$$

$$= \begin{bmatrix} I & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I & 0 \\ 0 & \cdots & 0 & \Lambda \end{bmatrix} \begin{bmatrix} I & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I & 0 \\ 0 & \cdots & 0 & \Lambda \end{bmatrix}^T,$$

where $\Lambda \in \mathbb{R}^{n_r \times l}$ is of full column rank. Therefore, from (5.14) there exists an orthogonal matrix Π such that

$$W_e = \begin{bmatrix} U_1 & U_2 & \cdots & U_r \Lambda \end{bmatrix} \Pi.$$

Moreover, the number of columns of Λ must equal $k - \sum_{i=1}^{r-1} n_i$ because of

$$W_e^T W_e = \Pi^T \begin{bmatrix} I_{n_1} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & I_{n_{r-1}} & 0 \\ 0 & \cdots & 0 & \Lambda^T \Lambda \end{bmatrix} \Pi$$

and

$$\operatorname{rank} W_e^T W_e = k.$$

The proof is completed by noting from (ii) of Lemma 3.1 that $\Lambda^T \Lambda = I$ if $r < p + 1$. □

Remark 5.2. In Theorem 5.1, $r = p + 1$ is equivalent to the fact that the kth largest eigenvalue of C is zero, which is true if C is nonsingular.

Another important remark concerning topological properties of the set of equilibria is in order.

Remark 5.3. Consider the case where the kth largest eigenvalue of C is nonzero. It is quite obvious from the above theorem that $\mathfrak{E}_s$ is a compact set in $\mathbb{R}^{n\times k}$. On the other hand, it is seen from the definition (5.4) that $\mathfrak{E}_s$ is an open subset of $\mathfrak{E}$. This means that $\mathfrak{E}_u$ is also a closed set in $\mathbb{R}^{n\times k}$ since the equilibrium set $\mathfrak{E}$ is a closed set. As a result, $\mathfrak{E}$ is divided into two disjoint nonempty closed sets $\mathfrak{E}_s$ and $\mathfrak{E}_u$ with the former being compact, which means that $\mathfrak{E}$ is disconnected.

Before tackling the stability issue, we need one more technical lemma below, which can be easily proved.

LEMMA 5.2. *Let $A \in \mathbb{R}^{n\times m}$. Then for any $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\left| \left\| (A+\Delta)^\dagger \right\| - \left\| A^\dagger \right\| \right| < \epsilon$$

provided

$$\|\Delta\| < \delta \quad \textit{and} \quad \operatorname{rank}(A+\Delta) = \operatorname{rank} A.$$

THEOREM 5.2. *Let $W_e \in \mathbb{R}^{n\times k}$ be an equilibrium of the ODE* (5.1). *For any given number $\epsilon > 0$, there exists a number $\delta > 0$ such that the solution $W(t)$ starting with $W(0) = W_e + \Delta W$ satisfies*

$$\|W(t) - W_e\| < \epsilon,$$

provided the perturbation ΔW obeys

$$\|\Delta W\| < \delta \tag{5.15}$$

and

$$\operatorname{rank}(W_e+\Delta W)^T \begin{bmatrix} U_1 & \cdots & U_i \end{bmatrix} = \operatorname{rank} W_e^T \begin{bmatrix} U_1 & \cdots & U_i \end{bmatrix}, \quad i = 1, \ldots, p+1. \tag{5.16}$$

Proof. Let $W(t)$ be the solution of (5.1) with $W(0) = W_e + \Delta W$ and put

$$F(t) \triangleq \exp(Ct)W(0)[I_k - W(0)^T W(0) + W(0)^T \exp(2Ct)W(0)]^{-1/2}.$$

Then by Lemma 3.2, there holds

$$W(t)W^T(t) = F(t)F^T(t), \quad \forall\, t \geq 0.$$

By Lemma 5.2, there exist two constants $\delta_1 > 0$ and $K_1 > 0$ such that

$$\left\| \left\{ \begin{bmatrix} U_1 & \cdots & U_i \end{bmatrix}^T (W_e + \Delta W) \right\}^\dagger \right\| < K_1, \quad i = 1, \ldots, p+1$$

whenever $\Delta W \in S(\delta_1)$, where $S(\delta)$ denotes the set of all the ΔW which satisfy (5.15) and (5.16). Consequently, by examining the derivation of (3.14) with particular

reference to (3.11), (3.12), and (3.13), it is seen that there is a constant $K_2 > 0$ such that

$$\|F(t) - \Theta\| \leq K_2 e^{-\mu t} \tag{5.17}$$

whenever $\Delta W \in S(\delta_1)$, where $\mu > 0$ and Θ are defined as in Theorem 3.1. This in turn implies the existence of a constant $K_3 > 0$ such that

$$\left\|[I - W(t)W^T(t)]CW(t)\right\| \leq K_3 e^{-\mu t}$$

whenever $\Delta W \in S(\delta_1)$. In this way, it is deduced that there exists $T > 0$ such that

$$\left\| \int_{t_1}^{t_2} [I - W(t)W^T(t)]CW(t)\, dt \right\| < \epsilon/3 \tag{5.18}$$

whenever $t_2 > t_1 \geq T$ and $\Delta W \in S(\delta_1)$. Furthermore, since the solution $W(t)$ is uniformly continuous with respect to the initial point $W(0)$ on any compact interval, there exists $0 < \delta < \min(\delta_1,\ \epsilon/3)$ such that for all $t \in [0,\ T]$,

$$\left\|[I - W(t)W^T(t)]CW(t) - (I - W_e W_e^T)CW_e\right\| < \frac{\epsilon}{2T},$$

i.e.,

$$\left\|[I - W(t)W^T(t)]CW(t)\right\| < \frac{\epsilon}{3T}$$

whenever $\|\Delta W\| \in S(\delta)$. This leads to

$$\left\| \int_0^T [I - W(t)W^T(t)]CW(t)\, dt \right\| < \epsilon/3 \tag{5.19}$$

whenever $\|\Delta W\| \in S(\delta)$. Combining (5.18) and (5.19) gives

$$\left\| \int_0^t [I - W(t)W^T(t)]CW(t)\, dt \right\| < 2\epsilon/3, \quad t \geq 0,$$

whenever $\|\Delta W\| \in S(\delta)$. Therefore, it is concluded that

$$\begin{aligned} \|W(t) - W_e\| &= \left\| \Delta W + \int_0^t [I - W(\tau)W^T(\tau)]CW(\tau)\, d\tau \right\| \\ &\leq \epsilon/3 + 2\epsilon/3 = \epsilon, \quad t \geq 0, \end{aligned}$$

provided $\|\Delta W\| \in S(\delta)$. The proof is completed. □

THEOREM 5.3. *Let $\mathfrak{E}_s$ be defined as in* (5.4). *Assume that the kth largest eigenvalue of the covariance matrix C is nonzero. Then, $W_e \in \mathbb{R}^{n\times k}$ is a stable equilibrium of the ODE* (5.1) *if and only if $W_e \in \mathfrak{E}_s$.*

Proof. Assuming that $W_e \in \mathfrak{E}_u$, we shall prove that W_e is an unstable equilibrium of the ODE (5.1). To do this, note that the distance of W_e from the closed set $\mathfrak{E}_s$ is positive. That is, one has

$$\epsilon_0 \triangleq \inf\{\|W - W_e\|\,;\ W \in \mathfrak{E}_s\} > 0.$$

On the other hand, for any number $\delta > 0$ there exists $W_0 \in \mathbb{R}^{n\times k}$ with $\|W_0 - W_e\| < \delta$ such that (5.6) and (5.7) are satisfied by $W = W_0$. By Lemma 5.1, the limiting solution W_∞ of (5.1) resulting from the initial condition $W(0) = W_0$ must be in $\mathfrak{E}_s$, which leads to

$$\|W_\infty - W_e\| \geq \epsilon_0.$$

This means that W_e is not a stable equilibrium. Thus, the "only if" part is proved.

Now assume that $W_e \in \mathfrak{E}_s$. Then (5.6) and (5.7) are satisfied with $W = W_e$. It is not difficult to see that this implies the existence of a number $\delta > 0$ such that the rank conditions

$$\text{rank}\begin{bmatrix} U_1 & \cdots & U_i \end{bmatrix}^T W = \text{rank}\begin{bmatrix} U_1 & \cdots & U_i \end{bmatrix}^T W_e, \quad i = 1, \ldots, p+1 \tag{5.20}$$

hold simultaneously for all W with $\|W - W_e\| < \delta$. By directly applying Theorem 5.2, it is shown that W_e is a stable equilibrium. □

The following result is an immediate consequence of Theorems 5.1 and 5.3 and stresses the relevance of the number of used neurons to principal subspace analysis.

COROLLARY 5.1. *The range of every stable equilibrium is a dominant eigenspace if and only if the kth largest eigenvalue of C is strictly larger than the $(k+1)$th largest one.*

6. Conclusions. This paper has studied the differential equation approximating Oja's subspace algorithm. A number of deep results have been obtained. The stability results are probably most important among them. More specifically, we have derived an explicit global exponential convergence rate for the equation in terms of the positive eigenvalues of the covariance matrix. The larger the positive eigenvalues and the deviations between any two of them, the greater the convergence rate. Given a generic starting point, the range of the limiting solution to the subspace equation is either a dominant eigenspace or a direct sum of a dominant eigenspace and a proper nonzero subspace of an eigenspace, depending on whether the kth largest eigenvalue is greater than the $(k+1)$th largest one, where k is the number of neurons used. Finally, all the stable equilibria have been found and parameterized in an explicit way. The solution to the subspace equation has been shown to converge to a stable equilibrium for almost all starting points.

Appendix.

Proof of Lemma 2.2. Let a singular value decomposition of X be

$$X = U\begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} V^T, \quad U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix},$$

where D is a positive definite diagonal matrix of the same dimension as U_{11}. Then one has

$$\begin{aligned}
&[I + X^T\Delta^{-1}X]^{-1} - \left[I - X^T\left(XX^T\right)^\dagger X\right] \\
&= X^T\left[\left(XX^T + \Delta\right)^{-1} - \left(XX^T\right)^\dagger\right]X \\
&= V\begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}\left(\begin{bmatrix} D^2 + \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}^{-1} - \begin{bmatrix} D^{-2} & 0 \\ 0 & 0 \end{bmatrix}\right)\begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}V^T \\
&= V\begin{bmatrix} D\left[\left(D^2 + \Phi\right)^{-1} - D^{-2}\right]D & 0 \\ 0 & 0 \end{bmatrix}V^T \\
&= -V\begin{bmatrix} D^{-1}\Phi D^{-1}\left(I + D^{-1}\Phi D^{-1}\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix}V^T, \qquad \text{(A.1)}
\end{aligned}$$

where

$$\Phi \triangleq \left\{ \begin{bmatrix} I & 0 \end{bmatrix} U^T \Delta^{-1} U \begin{bmatrix} I \\ 0 \end{bmatrix} \right\}^{-1}.$$

Since

$$\|\Phi\| \le \|\Delta\|,$$

it follows that

$$\left\| [I + X^T \Delta^{-1} X]^{-1} - \left[I - X^T \left(X X^T \right)^\dagger X \right] \right\| \le \left\| D^{-1} \right\|^2 \|\Delta\|,$$

which is desired due to $\left\| X^\dagger \right\| = \left\| D^{-1} \right\|$. □

Proof of Lemma 2.3. Set

$$X \triangleq I + X_1^T \Delta_1^2 X_1 + X_2^T \Delta_2^2 X_2 \quad \text{and} \quad Y \triangleq I + \Delta_1 X_1 X_1^T \Delta_1.$$

Then it is straightforward to check that

$$\Delta_1 X_1 X^{-1/2} = Y^{-1} \Delta_1 X_1 X^{1/2} - Y^{-1} \Delta_1 X_1 X_2^T \Delta_2^2 X_2 X^{-1/2}, \tag{A.2}$$

$$\Delta_1 X_1 \left(I + X_1^T \Delta_1^2 X_1 \right)^{-1/2} = Y^{-1} \Delta_1 X_1 \left(I + X_1^T \Delta_1^2 X_1 \right)^{1/2}. \tag{A.3}$$

Using the identity

$$X_1 = X_1 X_1^T X_1 \left(X_1^T X_1 \right)^\dagger,$$

one obtains

$$\begin{aligned} \left\| Y^{-1} \Delta_1 X_1 \right\| &= \left\| Y^{-1} \Delta_1 X_1 X_1^T \Delta_1 \Delta_1^{-1} X_1 \left(X_1^T X_1 \right)^\dagger \right\| \\ &= \left\| (I - Y^{-1}) \Delta_1^{-1} X_1 \left(X_1^T X_1 \right)^\dagger \right\| \\ &\le \left\| X_1^\dagger \right\| / \alpha_1. \end{aligned} \tag{A.4}$$

This leads to

$$\begin{aligned} & \left\| Y^{-1} \Delta_1 X_1 X_2^T \Delta_2^2 X_2 X^{-1/2} \right\| \\ \le\ & \left\| Y^{-1} \Delta_1 X_1 \right\| \left\| X_2^T \Delta_2 \right\| \left\| \Delta_2 X_2 X^{-1/2} \right\| \\ \le\ & \left\| X_1^\dagger \right\| \|X_2\| \, \alpha_2 / \alpha_1 \end{aligned} \tag{A.5}$$

as $\Delta_2 X_2 X^{-1} X_2^T \Delta_2 \le I$. On the other hand, note from (ii) of Lemma 2.1 that

$$\left\| X^{1/2} - \left(I + X_1^T \Delta_1^2 X_1 \right)^{1/2} \right\| \le \left\| X_2^T \Delta_2^2 X_2 \right\|^{1/2} \le \|X_2\| \, \alpha_2.$$

Thus, it follows from (A.2)–(A.5) that

$$\begin{aligned} & \left\| \Delta_1 X_1 \left[X^{-1/2} - \left(I + X_1^T \Delta_1^2 X_1 \right)^{-1/2} \right] \right\| \\ & \le \left\| Y^{-1} \Delta_1 X_1 \right\| \left\| X^{1/2} - \left(I + X_1^T \Delta_1^2 X_1 \right)^{1/2} \right\| + \left\| Y^{-1} \Delta_1 X_1 X_2^T \Delta_2^2 X_2 X^{-1/2} \right\| \\ & \le 2 \left\| X_1^\dagger \right\| \|X_2\| \, \alpha_2 / \alpha_1, \end{aligned}$$

which completes the proof. □

Proof of Lemma 2.4. Let $t > 0$ be fixed. With

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad \text{and} \quad C = \begin{bmatrix} C_1 & 0 \\ 0 & c_2 I \end{bmatrix},$$

one has

$$\begin{aligned} &\left\{I + X_1^T[\exp(2C_1t) - I]X_1 + [\exp(2c_2t) - 1]X_2^T X_2\right\}^{-1} \\ &= \left[I + X^T(\exp(2Ct) - I)X\right]^{-1} \\ &= I - X^T\left[XX^T + (\exp(2Ct) - I)^{-1}\right]^{-1} X \\ &= X^T\left\{\left(XX^T\right)^\dagger - \left[XX^T + (\exp(2Ct) - I)^{-1}\right]^{-1}\right\} X + I - X^T\left(XX^T\right)^\dagger X. \end{aligned}$$

Owing to the identity

$$X\left[I - X^T\left(XX^T\right)^\dagger X\right] = 0,$$

it follows from (i) of Lemma 2.1 that

$$\begin{aligned} &X_2\left\{I + X_1^T[\exp(2C_1t) - I]X_1 + [\exp(2c_2t) - 1]X_2^T X_2\right\}^{-1/2} \\ &= X_2\left\{X^T\left\{\left(XX^T\right)^\dagger - \left[XX^T + (\exp(2Ct) - I)^{-1}\right]^{-1}\right\} X\right\}^{1/2}. \end{aligned} \tag{A.6}$$

Choose two appropriate orthogonal matrices U, V and have the following partitions

$$X = U\begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} V^T, \quad U = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix},$$

where L is a positive definite diagonal matrix and U_{11} has the same number of rows as X_1 with the number of columns equal to the rank of X. Then one obtains

$$\begin{aligned} &X^T\left\{\left(XX^T\right)^\dagger - \left[XX^T + (\exp(2Ct) - I)^{-1}\right]^{-1}\right\} X \\ &= V\begin{bmatrix} L^{-1}\Delta\left(L^2 + \Delta\right)^{-1} L & 0 \\ 0 & 0 \end{bmatrix} V^T, \end{aligned} \tag{A.7}$$

where

$$\begin{aligned} \Delta &\triangleq \left\{\begin{bmatrix} U_{11}^T & U_{21}^T \end{bmatrix}[\exp(2Ct) - I]\begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix}\right\}^{-1} \\ &= \left\{U_{11}^T[\exp(2C_1t) - I]U_{11} + [\exp(2c_2t) - 1]U_{21}^T U_{21}\right\}^{-1} \\ &= \left\{U_{11}^T[\exp(2C_1t) - \exp(2c_2t)I]U_{11} + [\exp(2c_2t) - 1]I\right\}^{-1} \\ &= \left(e^{2c_2t} - 1\right)^{-1}\left[I + U_{11}^T\frac{\exp(2C_1t) - \exp(2c_2t)I}{\exp(2c_2t) - 1}U_{11}\right]^{-1}. \end{aligned}$$

As

$$\frac{\exp(2C_1t) - \exp(2c_2t)I}{\exp(2c_2t) - 1} \geq \frac{c_1 - c_2}{c_1}e^{2(c_1 - c_2)t}$$

from Lemma 2.2 there results

$$(e^{2c_2t}-1)\Delta - Q \le \beta^2 \left\|U_{11}^{\dagger}\right\|^2 e^{-2(c_1-c_2)t}, \tag{A.8}$$

where $Q \triangleq [I - U_{11}^T \left(U_{11}U_{11}^T\right)^{\dagger} U_{11}] \le I$. Now with (A.8) and

$$\begin{aligned}\left\|\left(I + L^{-1}\Delta L^{-1}\right)^{-1} - I\right\| &= \left\|\left(I + L\Delta^{-1}L\right)^{-1}\right\| \\ &\le \left\|\left[I + \left(e^{2c_2t}-1\right)L^2\right]^{-1}\right\| \\ &\le \left(1 + \frac{e^{2c_2t}-1}{\|L^{-2}\|}\right)^{-1} \\ &\le \left(1 + \left\|L^{-2}\right\|\right) e^{-2c_2t},\end{aligned}$$

we are led to

$$\begin{aligned}&\left\|e^{2c_2t}L^{-1}\Delta L^{-1}\left(I + L^{-1}\Delta L^{-1}\right)^{-1} - L^{-1}QL^{-1}\right\| \\ &= \left\|L^{-1}\left[(e^{2c_2t}-1)\Delta - Q\right]L^{-1}\left(I + L^{-1}\Delta L^{-1}\right)^{-1}\right. \\ &\qquad \left. + \left(L^{-1}QL^{-1} - I\right)\left[\left(I + L^{-1}\Delta L^{-1}\right)^{-1} - I\right]\right\| \\ &\le \beta^2 \left\|L^{-1}\right\|^2 \left\|U_{11}^{\dagger}\right\|^2 e^{-2(c_1-c_2)t} + \left\|L^{-1}QL^{-1} - I\right\| \left\|\left(I + L^{-1}\Delta L^{-1}\right)^{-1} - I\right\| \\ &\le \beta^2 \left\|L^{-1}\right\|^2 \left\|U_{11}^{\dagger}\right\|^2 e^{-2(c_1-c_2)t} + \left(1 + \left\|L^{-2}\right\|\right)^2 e^{-2c_2t}. \qquad \text{(A.9)}\end{aligned}$$

Since

$$X_1 = \begin{bmatrix}U_{11} & U_{12}\end{bmatrix}\begin{bmatrix}L & 0\\ 0 & 0\end{bmatrix}V^T, \tag{A.10}$$

it is evident that

$$X_1X_1^T = U_{11}L^2U_{11}^T \le \|L\|^2 U_{11}U_{11}^T$$

and that $X_1X_1^T$ has the same null space as $U_{11}U_{11}^T$. As such, the minimum nonzero singular value of U_{11} is no less than that of $X_1X_1^T$ divided by $\|L\|$, i.e.,

$$\left\|U_{11}^{\dagger}\right\| \le \|L\| \left\|X_1^{\dagger}\right\|.$$

By noting that

$$\|L\| = \|X\| \quad \text{and} \quad \left\|L^{-1}\right\| = \left\|X^{\dagger}\right\|,$$

(A.9) results in

$$\begin{aligned}&\left\|e^{2c_2t}L^{-1}\Delta L^{-1}\left(I + L^{-1}\Delta L^{-1}\right)^{-1} - L^{-1}QL^{-1}\right\| \\ &\le \left[\beta^2 \|X\|^2 \left\|X^{\dagger}\right\|^2 \left\|X_1^{\dagger}\right\|^2 + \left(1 + \left\|L^{-2}\right\|\right)^2\right] e^{-2\alpha t}. \qquad \text{(A.11)}\end{aligned}$$

On the other hand, in view of (A.10) and

$$S = V\begin{bmatrix}L^{-2} & 0\\ 0 & 0\end{bmatrix}V^T,$$

it is readily checked that

$$S - SX_1^T \left(X_1 S X_1^T\right)^\dagger X_1 S = V \begin{bmatrix} L^{-1}QL^{-1} & 0 \\ 0 & 0 \end{bmatrix} V^T.$$

Consequently, it follows from (A.6), (A.7), and (A.11) that

$$\begin{aligned}
&\left\| e^{c_2 t} X_2 \left\{ I + X_1^T [\exp(2C_1 t) - I] X_1 + (e^{2c_2 t} - 1) X_2^T X_2 \right\}^{-1/2} \right. \\
&\qquad \left. - X_2 \left[S - SX_1^T \left(X_1 S X_1^T\right)^\dagger X_1 S \right]^{1/2} \right\| \\
&= \left\| X_2 V \left(\begin{bmatrix} e^{2c_2 t} L^{-1} \Delta L^{-1} \left(I + L^{-1} \Delta L^{-1}\right)^{-1} & 0 \\ 0 & 0 \end{bmatrix}^{1/2} \right. \right. \\
&\qquad \left. \left. - \begin{bmatrix} L^{-1}QL^{-1} & 0 \\ 0 & 0 \end{bmatrix}^{1/2} \right) V^T \right\| \\
&\le \|X_2\| \left\| e^{2c_2 t} L^{-1} \Delta L^{-1} \left(I + L^{-1} \Delta L^{-1}\right)^{-1} - L^{-1} Q L^{-1} \right\|^{1/2} \\
&\le \|X_2\| \sqrt{\beta^2 \|X\|^2 \|X^\dagger\|^2 \left\| X_1^\dagger \right\|^2 + \left(1 + \|X^\dagger\|^2\right)^2}\, e^{-\alpha t} \\
&\le \|X_2\| \left(\beta \|X\| \left\| X^\dagger \right\| \left\| X_1^\dagger \right\| + 1 + \left\| X^\dagger \right\|^2 \right) e^{-\alpha t}
\end{aligned}$$

as required. □

Proof of Lemma 3.1. From the identity

$$\ker \left[I - X^T \left(X X^T\right)^\dagger X \right] = \operatorname{range} X^T, \tag{A.12}$$

it is true that $V_i^2 \mathcal{U}_{i-1}^T = 0$. But, V_i is symmetric; hence, it follows that $V_i \mathcal{U}_{i-1}^T = 0$, i.e., $\operatorname{range} \mathcal{U}_{i-1}^T \subset \ker V_i$. That $\ker \mathcal{U}_i \subset \ker V_i$ can be seen from

$$\ker \mathcal{U}_i \subset \ker \left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger$$

and

$$\ker \left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \subset \ker V_i^2 = \ker V_i.$$

So there results

$$\operatorname{range}\, \mathcal{U}_{i-1}^T + \ker \mathcal{U}_i \subset \ker V_i. \tag{A.13}$$

To prove the reverse inclusion, let

$$x \in \ker V_i \quad \text{and} \quad y = \left[\left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \right]^{1/2} x.$$

Then there holds

$$y^T \left[I - \left[\left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \right]^{1/2} \mathcal{U}_{i-1}^T \left[\mathcal{U}_{i-1} \left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \mathcal{U}_{i-1}^T \right]^\dagger \mathcal{U}_{i-1} \left[\left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \right]^{1/2} \right] y = 0,$$

or equivalently,

$$y \in \ker \left\{ I - \left[\left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \right]^{1/2} \mathcal{U}_{i-1}^T \left[\mathcal{U}_{i-1} \left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \mathcal{U}_{i-1}^T \right]^\dagger \mathcal{U}_{i-1} \left[\left(\mathcal{U}_i^T \mathcal{U}_i\right)^\dagger \right]^{1/2} \right\}.$$

Again from the identity (A.12), it is deduced that

$$y \in \text{range}\left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2}\mathcal{U}_{i-1}^T.$$

That is, there is some vector z such that

$$y = \left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2}\mathcal{U}_{i-1}^T z$$

leading to

$$\left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2}\left(x - \mathcal{U}_{i-1}^T z\right) = 0.$$

Thus, it is concluded that

$$x - \mathcal{U}_{i-1}^T z \in \ker \mathcal{U}_i.$$

Since x is arbitrary, it is verified that

$$\ker V_i \subset \text{range}\,\mathcal{U}_{i-1}^T + \ker \mathcal{U}_i.$$

Combining this with (A.13) and noting that $\text{range}\,\mathcal{U}_{i-1}^T \cap \ker \mathcal{U}_i = \emptyset$ yields (i).

To prove (ii), note that for $i < p+1$,

$$(U_i^T W_0)^T U_i^T W_0 = \mathcal{U}_i^T\mathcal{U}_i - \mathcal{U}_{i-1}^T\mathcal{U}_{i-1}, \quad V_i^2\mathcal{U}_i^T\mathcal{U}_i V_i^2 = V_i^2,$$

which, together with (i), implies that

$$V_i(U_i^T W_0 V_i)^T(U_i^T W_0 V_i)V_i = V_i^2\left(\mathcal{U}_i^T\mathcal{U}_i - \mathcal{U}_{i-1}^T\mathcal{U}_{i-1}\right)V_i^2 = V_i^2.$$

Pre- and postmultiplying the above by $V_i^\dagger$ gives rise to (ii).

Since V_i is symmetric, (iii) is equivalent to the fact that $V_i V_j = 0$ for $i < j$. Let us assume that i and j are such that $i < j$. Then, all the rows of $\mathcal{U}_i$ are the rows of $\mathcal{U}_{j-1}$. Further, note that with

$$Q = I - \left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2}\mathcal{U}_{i-1}^T\left[\mathcal{U}_{i-1}\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\mathcal{U}_{i-1}^T\right]^\dagger\mathcal{U}_{i-1}\left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2},$$

V_i can be expressed as

$$\begin{aligned} V_i &= \left\{\left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2} Q\left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2}\right\}^{1/2} \\ &= \left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2} Q\left\{\left[Q\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger Q\right]^{1/2}\right\}^\dagger Q\left[\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right]^{1/2}. \end{aligned} \tag{A.14}$$

Here, the fact that $Q^2 = Q$ has been used. On the other hand, it is clear from (i) that $\mathcal{U}_i V_j^2 = 0$, implying that

$$\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger V_j = 0$$

due to

$$\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger = \mathcal{U}_i^T\left[\left(\mathcal{U}_i\mathcal{U}_i^T\right)^\dagger\right]^2\mathcal{U}_i.$$

In this way, it is seen that $[(\mathcal{U}_i^T\mathcal{U}_i)^\dagger]^{1/2}V_j = 0$. This together with (A.14) gives $V_iV_j = 0$.

To prove (iv), assume that $\operatorname{rank}\mathcal{U}_i \neq n_i + \operatorname{rank}\mathcal{U}_{i-1}$, which implies the existence of some nonzero $x \in \operatorname{range}\mathcal{U}_{i-1}^T \cap \operatorname{range} W_0^TU_i$. By (i), one has $x \in \ker V_i$. Combining this with $x \in \operatorname{range} W_0^TU_i$ yields that $\ker V_iW_0^TU_i \neq \emptyset$. This means that $U_i^TW_0V_i^2W_0^TU_i$ is not even of full rank. Now assume that $\operatorname{rank}\mathcal{U}_i = n_i + \operatorname{rank}\mathcal{U}_{i-1}$, which implies that $U_i^TW_0$ is of full row rank. Let $\mathcal{U}_{i-1} = PQ$, where P is of full column rank and Q is of full row rank. Then, there holds

$$\mathcal{U}_i = \begin{bmatrix} \mathcal{U}_{i-1} \\ U_i^TW_0 \end{bmatrix} = \begin{bmatrix} P & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} Q \\ U_i^TW_0 \end{bmatrix},$$

where the two factor matrices are of full column rank and full row rank, respectively. Thus, it can be verified that

$$\mathcal{U}_i \left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger \mathcal{U}_i^T = \begin{bmatrix} P\left(P^TP\right)^{-1}P^T & 0 \\ 0 & I \end{bmatrix}. \tag{A.15}$$

In particular, one has

$$U_i^TW_0\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger W_0^TU_i = I \quad \text{and} \quad U_i^TW_0\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger \mathcal{U}_{i-1}^T = 0.$$

As a consequence, it is concluded that

$$\begin{aligned} &U_i^TW_0V_i^2W_0^TU_i \\ &= U_i^TW_0\left\{\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger - \left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger \mathcal{U}_{i-1}^T\left[\mathcal{U}_{i-1}\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger \mathcal{U}_{i-1}^T\right]^\dagger \mathcal{U}_{i-1}\left(\mathcal{U}_i^T\mathcal{U}_i\right)^\dagger\right\} W_0^TU_i \\ &= I. \qquad \square \end{aligned}$$

Acknowledgments. The author is grateful for corrections and useful comments by the reviewers.

REFERENCES

[1] R. Williams, *Feature Discovery through Error-correcting Learning*, Tech. report 8501, Institute of Cognitive Science, University of California at San Diego, 1985.

[2] P. Baldi, *Linear learning: Landscapes and algorithms*, in Advances in Neural Information Processing Systems 1, D. S. Touretzky, ed., Morgan-Kaufmann, San Francisco, CA, 1988.

[3] E. Oja, *Neural networks, principal components, and subspaces*, Internat. J. Neural Systems, 1 (1989), pp. 61–68.

[4] L. Xu, A. Krzyzak, and E. Oja, *Neural nets for dual subspace pattern recognition*, Internat. J. Neural Systems, 2 (1991), pp. 169–187.

[5] P. Baldi and K. Hornik, *Learning in linear neural networks: A survey*, IEEE Trans. Neural Networks, 6 (1995), pp. 837–858.

[6] E. Oja, *A simplified neuron model as a principal component analyzer*, J. Math. Biol., 15 (1982), pp. 267–273.

[7] E. Oja and J. Karhunen, *On stochastic approximation of eigenvectors and eigenvalues of the expectation of a random matrix*, J. Math. Anal. Appl., 106 (1985), pp. 69–84.

[8] K. Hornik and C. M. Kuan, *Convergence analysis of local feature extraction algorithms*, Neural Networks, 5 (1992), pp. 229–240.

[9] A. Krogh and J. Hertz, *Hebbian learning of principal components*, in Parallel Processing in Neural Systems and Computers, R. Eckmiller, G. Hartmann, and G. Hauske, eds., Elsevier, North–Holland, Amsterdam, 1990, pp. 183–186.

[10] W.-Y. Yan, U. Helmke, and J. B. Moore, *Global analysis of Oja's flow for neural networks*, IEEE Trans. Neural Networks, 5 (1994), pp. 674–683.
[11] L. Xu, *Least mean square error reconstruction principle for self-organizing neural nets*, Neural Networks, 6 (1993), pp. 627–648.
[12] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*, Springer-Verlag, London, 1994.
[13] E. Oja, *Principal components, minor components, and linear neural networks*, Neural Networks, 5 (1992), pp. 927–935.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 956–982, October 1998

009

RELATIVE PERTURBATION THEORY: I. EIGENVALUE AND SINGULAR VALUE VARIATIONS*

REN-CANG LI†

Abstract. The classical perturbation theory for Hermitian matrix eigenvalue and singular value problems provides bounds on the absolute differences between approximate eigenvalues (singular values) and the true eigenvalues (singular values) of a matrix. These bounds may be bad news for small eigenvalues (singular values), which thereby suffer worse relative uncertainty than large ones. However, there are situations where even small eigenvalues are determined to high relative accuracy by the data much more accurately than the classical perturbation theory would indicate. In this paper, we study how eigenvalues of a Hermitian matrix A change when it is perturbed to $\widetilde{A} = D^*AD$, where D is close to a unitary matrix, and how singular values of a (nonsquare) matrix B change when it is perturbed to $\widetilde{B} = D_1^*BD_2$, where D_1 and D_2 are nearly unitary. It is proved that under these kinds of perturbations small eigenvalues (singular values) suffer relative changes no worse than large eigenvalues (singular values). Many well-known perturbation theorems, including the Hoffman–Wielandt and Weyl–Lidskii theorems, are extended.

Key words. multiplicative perturbation, relative perturbation theory, relative distance, eigenvalue, singular value, graded matrix

AMS subject classifications. 15A18, 15A42, 65F15, 65F35, 65G99

PII. S089547989629849X

1. Introduction.

The classical perturbation theory for Hermitian matrix eigenvalue problems provides bounds on the absolute differences $|\lambda - \widetilde{\lambda}|$ between approximate eigenvalues $\widetilde{\lambda}$ and the true eigenvalues λ of a Hermitian matrix A. When $\widetilde{\lambda}$ is computed using standard numerical software, the bounds on $|\lambda - \widetilde{\lambda}|$ are typically only moderately bigger than $\epsilon\|A\|$ [15, 33, 40], where ϵ is the rounding error threshold characteristic of the computer's arithmetic. These bounds are bad news for small eigenvalues, which thereby suffer worse relative uncertainty than large ones.

Generally, the classical error bounds are best possible if perturbations are arbitrary. However, there are situations where perturbations have special structures and, under these special perturbations, even small eigenvalues (singular values) are determined to high relative accuracy by the data much more accurately than the classical perturbation theory would indicate. A relative perturbation theory is then called for to exploit the situations for better bounds on the *relative* differences between $\widetilde{\lambda}$ and λ.

*Received by the editors February 2, 1996; accepted for publication (in revised form) by R. Bhatia June 23, 1997; published electronically July 7, 1998. A preliminary version of this paper appeared as Technical Report UCB//CSD-94-855, Computer Science Division, Department of EECS, University of California at Berkeley, 1994, and also appeared as LAPACK working note 85 (revised January 1996) available online at http://www.netlib.org/lapack/lawns/lawns84.ps. This research was supported in part by Argonne National Laboratory under grant 20552402, by the University of Tennessee through the Advanced Research Projects Agency under contract DAAL03-91-C-0047, by the National Science Foundation under grant ASC-9005933, by National Science Infrastructure grants CDA-8722788 and CDA-9401156, and by a Householder Fellowship in Scientific Computing at Oak Ridge National Laboratory, supported by the Applied Mathematical Sciences Research Program, Office of Energy Research, United States Department of Energy contract DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.

http://www.siam.org/journals/simax/19-4/29849.html

†Mathematical Science Section, Oak Ridge National Laboratory, P.O. Box 2008, Bldg. 6012, Oak Ridge, TN 37831-6367. Present address: Department of Mathematics, University of Kentucky, Lexington, KY 40506 (rcli@cs.uky.edu).

The development of such a theory goes back to Kahan [20] and is becoming a very active area of research [1, 6, 7, 8, 9, 11, 12, 14, 16, 10, 28, 34]. In this paper, we develop a theory by a unifying treatment that sharpens some existing bounds and covers many previously studied cases. We shall deal with perturbations that have multiplicative structures; namely, perturbations to unperturbed matrices are realized by multiplying the unperturbed ones with matrices that are nearly unitary. (To be exact, our theorems only require those multiplying matrices to be nonsingular, but our bounds are interesting only when they are close to some unitary matrices.) For Hermitian eigenvalue problems, we shall assume that A is perturbed to $\widetilde{A} = D^*AD$, where D is nonsingular; and for singular value problems we shall consider that B is perturbed to $\widetilde{B} = D_1^*BD_2$, where D_1 and D_2 are nonsingular. It is proved that these kinds of perturbations introduce no bigger uncertainty to small eigenvalues (in magnitude) and small singular values than they would to large ones. Although special, these perturbations cover *componentwise relative perturbations* of entries of symmetric tridiagonal matrices with zero diagonal [8, 20] and *componentwise relative perturbations* of entries of bidiagonal and biacyclic matrices [1, 7, 8]. More realistically, perturbations of graded nonnegative Hermitian matrices [9, 28] and perturbations of graded matrices of singular value problems [9, 28] can be transformed to take forms of multiplicative perturbations as will be seen from later proofs.

Additive perturbations are the most general in the sense that if A is perturbed to $\widetilde{A}$, the only possible known information is on some norm of $\Delta A \stackrel{\text{def}}{=} \widetilde{A} - A$. Such perturbations, no matter how small, may not guarantee relative accuracy in eigenvalues (singular values) of the matrix under consideration. For example, when A is singular, $\widetilde{A}$ can be made nonsingular no matter how small a norm of ΔA is; thus some zero eigenvalues are perturbed to nonzero ones and therefore lose their relative accuracy completely. (Retaining any relative accuracy of a zero at all ends up not changing it.)

The rest of this paper is organized as follows. Section 2 defines two kinds of relative distances ϱ_p $(1 \le p \le \infty)$ and χ, and Appendices A and B present proofs of some crucial properties of ϱ_p and χ needed in this paper. We devote two sections to present and discuss our main theorems—section 3 for relative perturbation theorems for Hermitian matrix eigenvalue problems and section 4 for relative perturbation theorems for singular value problems. Long proofs of our main theorems are postponed to sections 5 and 6. Section 7 briefly discusses how our relative perturbation theorems can be applied to generalized eigenvalue problems and generalized singular value problems.

Notation. We shall adopt the following convention: capital letters denote unperturbed matrices and capital letters with *tildes* denote their perturbed matrices. For example, X *is perturbed to* $\widetilde{X}$. Throughout the paper, capital letters are for matrices, lowercase Latin letters for column vectors or scalars, and lowercase Greek letters for scalars. Also,

$\mathbb{C}^{m\times n}$:	the set of $m \times n$ complex matrices, and $\mathbb{C}^m = \mathbb{C}^{m\times 1}$;
$\mathbb{R}^{m\times n}$:	the set of $m \times n$ real matrices, and $\mathbb{R}^m = \mathbb{R}^{m\times 1}$;
$\mathbb{U}_n$:	the set of $n \times n$ unitary matrices;
$0_{m,n}$:	the $m \times n$ zero matrix (we may simply write 0 instead);
I_n:	the $n \times n$ identity matrix (we may simply write I instead);

X^*: the conjugate transpose of a matrix X;
$\lambda(X)$: the set of the eigenvalues of X, counted according to their algebraic multiplicities;
$\sigma(X)$: the set of the singular values of X, counted according to their algebraic multiplicities;
$\sigma_{\min}(X)$: the smallest singular value of $X \in \mathbb{C}^{m\times n}$;
$\sigma_{\max}(X)$: the largest singular value of $X \in \mathbb{C}^{m\times n}$;
$\|X\|_2$: the spectral norm of X, i.e., $\sigma_{\max}(X)$;
$\|X\|_{\mathrm{F}}$: the Frobenius norm of X, i.e., $\sqrt{\sum_{i,j}|x_{ij}|^2}$, where $X = (x_{ij})$.

2. Relative distances. Classically, the relative error in $\widetilde{\alpha} = \alpha(1+\delta)$ as an approximation to α is measured by

$$\delta = \text{relative error in } \widetilde{\alpha} = \frac{\widetilde{\alpha}-\alpha}{\alpha}. \tag{2.1}$$

When $|\delta| \le \epsilon$, we say that the relative perturbation to α is at most ϵ (see, e.g., [8]). Such a measurement lacks mathematical properties upon which a nice relative perturbation theory can be built; for example, it lacks symmetry between α and $\widetilde{\alpha}$ and thus it cannot be a metric. Nonetheless, it is good enough and is convenient to use for measuring correct digits in numerical approximations.

Our new *relative distances* have better mathematical properties, such as symmetry in the arguments. Topologically they are all equivalent to the classical δ-measurement defined by (2.1). The *p-relative distance* between α, $\widetilde{\alpha} \in \mathbb{C}$ is defined as

$$\varrho_p(\alpha,\widetilde{\alpha}) \stackrel{\text{def}}{=} \frac{|\alpha-\widetilde{\alpha}|}{\sqrt[p]{|\alpha|^p+|\widetilde{\alpha}|^p}} \quad \text{for } 1 \le p \le \infty. \tag{2.2}$$

We define, for convenience, $0/0 \stackrel{\text{def}}{=} 0$. ϱ_∞ has been used by Deift et al. [6] to define relative gaps. Another *relative distance* that is of interest to us is

$$\chi(\alpha,\widetilde{\alpha}) \stackrel{\text{def}}{=} \frac{|\alpha-\widetilde{\alpha}|}{\sqrt{|\alpha\widetilde{\alpha}|}}. \tag{2.3}$$

This χ-distance has been used by Barlow and Demmel [1] and Demmel and Veselić [9] to define relative gaps between the spectra of two matrices.

Appendix B will show that ϱ_p $(1 \le p \le \infty)$ *is indeed a metric on* $\mathbb{R}$; see also Li [24]. (We suspect that ϱ_p is a metric on $\mathbb{C}$ also, but we cannot give a proof at this point.) Unfortunately χ violates the triangle inequality and thus cannot be a metric. In fact, one can prove that $\chi(\alpha,\gamma) > \chi(\alpha,\beta)+\chi(\beta,\gamma)$ for $\alpha < \beta < \gamma$; see Lemma 6.1.

We refer the reader to Li [24] for a detailed study of the two relative distances. Here, only properties that are most relevant to our relative perturbation theory will be presented, and those proofs that require little work and seem to be straightforward are omitted. Complicated proofs will be given in Appendix A.

PROPOSITION 2.1 (see [24]). *Let* α, $\widetilde{\alpha} \in \mathbb{R}$.

1. *For* $0 \le \epsilon < 1$,

$$\left|\frac{\widetilde{\alpha}}{\alpha}-1\right| \le \epsilon \Rightarrow \varrho_p(\alpha,\widetilde{\alpha}) \le \frac{\epsilon}{\sqrt[p]{1+(1-\epsilon)^p}}, \tag{2.4}$$

$$\left|\frac{\widetilde{\alpha}}{\alpha}-1\right| \le \epsilon \Rightarrow \chi(\alpha,\widetilde{\alpha}) \le \frac{\epsilon}{\sqrt{1-\epsilon}}. \tag{2.5}$$

2. *For* $0 \le \epsilon < 1$,

$$\varrho_p(\alpha, \widetilde{\alpha}) \le \epsilon \Rightarrow \max\left\{\left|\frac{\widetilde{\alpha}}{\alpha} - 1\right|, \left|\frac{\alpha}{\widetilde{\alpha}} - 1\right|\right\} \le \frac{2^{1/p}\,\epsilon}{1-\epsilon}. \tag{2.6}$$

For $0 \le \epsilon < 2$,

$$\chi(\alpha, \widetilde{\alpha}) \le \epsilon \Rightarrow \max\left\{\left|\frac{\widetilde{\alpha}}{\alpha} - 1\right|, \left|\frac{\alpha}{\widetilde{\alpha}} - 1\right|\right\} \le \left(\frac{\epsilon}{2} + \sqrt{1 + \frac{\epsilon^2}{4}}\right)\epsilon. \tag{2.7}$$

3. *Asymptotically,*

$$\lim_{\widetilde{\alpha}\to\alpha} \frac{\varrho_p(\alpha, \widetilde{\alpha})}{\left|\frac{\widetilde{\alpha}}{\alpha} - 1\right|} = 2^{1/p} \quad \textit{and} \quad \lim_{\widetilde{\alpha}\to\alpha} \frac{\chi(\alpha, \widetilde{\alpha})}{\left|\frac{\widetilde{\alpha}}{\alpha} - 1\right|} = 1.$$

Thus (2.4), (2.6), (2.5), *and* (2.7) *are at least asymptotically sharp.*

The following proposition establishes a relation between ϱ_p and χ.

PROPOSITION 2.2 (see [24]). *For* $\alpha, \widetilde{\alpha} \in \mathbb{C}$,

$$\varrho_p(\alpha, \widetilde{\alpha}) \le 2^{-1/p}\,\chi(\alpha, \widetilde{\alpha}),$$

and the equality holds if and only if $|\alpha| = |\widetilde{\alpha}|$.

Next we ask *what are the best one-one pairings between two sets of n real numbers*? Such a question will become important later in this paper when we try to pair the eigenvalues or the singular values of one matrix to those of another.

PROPOSITION 2.3 (see [24]). *Let* $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ *and* $\{\widetilde{\alpha}_1, \widetilde{\alpha}_2, \ldots, \widetilde{\alpha}_n\}$ *be two sets of n real numbers ordered in descending order, i.e.,*

$$\alpha_1 \ge \alpha_2 \ge \cdots \ge \alpha_n, \qquad \widetilde{\alpha}_1 \ge \widetilde{\alpha}_2 \ge \cdots \ge \widetilde{\alpha}_n. \tag{2.8}$$

We have for $p = 1$,

$$\max_{1\le i\le n} \varrho_1(\alpha_i, \widetilde{\alpha}_i) = \min_{\tau} \max_{1\le i\le n} \varrho_1(\alpha_i, \widetilde{\alpha}_{\tau(i)}).$$

For $p > 1$, *if in addition all* α_i*'s and* $\widetilde{\alpha}_j$*'s are nonnegative,*

$$\max_{1\le i\le n} \varrho_p(\alpha_i, \widetilde{\alpha}_i) = \min_{\tau} \max_{1\le i\le n} \varrho_p(\alpha_i, \widetilde{\alpha}_{\tau(i)}). \tag{2.9}$$

Both minimizations are taken over all permutations τ *of* $\{1, 2, \ldots, n\}$.

Proofs of this proposition and Proposition 2.4 below are given in Appendix A.

Remark 2.1. Equation (2.9) of Proposition 2.3 may fail if not all the α_i's and $\widetilde{\alpha}_j$'s are of the same sign. A *counterexample* is as follows: $n = 2$ and

$$\alpha_1 = 1 > \alpha_2 = -2 \quad \text{and} \quad \widetilde{\alpha}_1 = 4 > \widetilde{\alpha}_2 = 2.$$

Then for $p > 1$,

$$\begin{aligned} &\max\left\{\varrho_p(\alpha_1, \widetilde{\alpha}_1),\, \varrho_p(\alpha_2, \widetilde{\alpha}_2)\right\} = \varrho_p(\alpha_2, \widetilde{\alpha}_2) = 2^{1-1/p} \\ &> \frac{6}{\sqrt[p]{2^p + 4^p}} = \varrho_p(\alpha_2, \widetilde{\alpha}_1) = \max\left\{\varrho_p(\alpha_1, \widetilde{\alpha}_2),\, \varrho_p(\alpha_2, \widetilde{\alpha}_1)\right\}. \end{aligned}$$

Remark 2.2. Given two sets of α_i's and $\widetilde{\alpha}_j$'s ordered as in (2.8), generally,

$$\sum_{i=1}^{n} [\varrho_p(\alpha_i, \widetilde{\alpha}_i)]^2 \neq \min_{\tau} \sum_{i=1}^{n} \left[\varrho_p(\alpha_i, \widetilde{\alpha}_{\tau(i)})\right]^2, \tag{2.10}$$

even if all α_i, $\widetilde{\alpha}_j > 0$. Here is a *counterexample*: $n = 2$,

$$\widetilde{\alpha}_1 > \alpha_1 = \widetilde{\alpha}_1/2 > \widetilde{\alpha}_2 > \alpha_2 > 0,$$

where α_2 is sufficiently close to 0, and $\widetilde{\alpha}_2$ is sufficiently close to α_1 which is fixed. Since, as $\alpha_2 \to 0^+$ and $\widetilde{\alpha}_2 \to \alpha_1^-$,

$$\begin{aligned} &[\varrho_p(\alpha_1, \widetilde{\alpha}_2)]^2 + [\varrho_p(\alpha_2, \widetilde{\alpha}_1)]^2 \to 1, \\ &[\varrho_p(\alpha_1, \widetilde{\alpha}_1)]^2 + [\varrho_p(\alpha_2, \widetilde{\alpha}_2)]^2 \to \frac{1}{\sqrt[p]{2^p + 1}} + 1, \end{aligned}$$

(2.10) must fail for some $\widetilde{\alpha}_1 > \alpha_1 = \widetilde{\alpha}_1/2 > \widetilde{\alpha}_2 > \alpha_2 > 0$.

PROPOSITION 2.4 (see [24]). *Let* $\{\alpha_1, \ldots, \alpha_n\}$ *and* $\{\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_n\}$ *be two sets of* n *positive numbers ordered as in* (2.8). *Then*

$$\max_{1 \le i \le n} \chi(\alpha_i, \widetilde{\alpha}_i) = \min_{\tau} \max_{1 \le i \le n} \chi(\alpha_i, \widetilde{\alpha}_{\tau(i)}), \tag{2.11}$$

$$\sum_{i=1}^{n} [\chi(\alpha_i, \widetilde{\alpha}_i)]^2 = \min_{\tau} \sum_{i=1}^{n} \left[\chi(\alpha_i, \widetilde{\alpha}_{\tau(i)})\right]^2, \tag{2.12}$$

where the minimization is taken over all permutations τ *of* $\{1, 2, \ldots, n\}$.

Remark 2.3. Both (2.11) and (2.12) of Proposition 2.4 may fail if the α_i's and $\widetilde{\alpha}_j$'s are not all of the same sign. A *counterexample* for (2.11) is that $n = 2$ and

$$\alpha_1 = 1 > \alpha_2 = -1 \quad \text{and} \quad \widetilde{\alpha}_1 = 2 > \widetilde{\alpha}_2 = \frac{1}{4},$$

for which

$$\begin{aligned} &\max\{\chi(\alpha_1, \widetilde{\alpha}_1),\, \chi(\alpha_2, \widetilde{\alpha}_2)\} = \max\left\{1/\sqrt{2}, 5/2\right\} = 5/2 \\ &> 3/\sqrt{2} = \max\left\{3/2, 3/\sqrt{2}\right\} = \max\{\chi(\alpha_1, \widetilde{\alpha}_2),\, \chi(\alpha_2, \widetilde{\alpha}_1)\}. \end{aligned}$$

A *counterexample* for (2.12) is that $n = 2$ and

$$\alpha_1 = 1 > \alpha_2 = -2 \quad \text{and} \quad \widetilde{\alpha}_1 = 2 > \widetilde{\alpha}_2 = 1,$$

for which

$$\begin{aligned} [\chi(\alpha_1, \widetilde{\alpha}_1)]^2 + [\chi(\alpha_2, \widetilde{\alpha}_2)]^2 &= \left(1/\sqrt{2}\right)^2 + \left(3/\sqrt{2}\right)^2 = 5 \\ &> 4 = 0^2 + \left(4/\sqrt{4}\right)^2 = [\chi(\alpha_1, \widetilde{\alpha}_2)]^2 + [\chi(\alpha_2, \widetilde{\alpha}_1)]^2. \end{aligned}$$

3. Relative perturbation theorems for Hermitian matrix eigenvalue problems. Throughout the section, A, $\widetilde{A} \in \mathbb{C}^{n\times n}$ are Hermitian and one is a perturbation of the other. Denote their eigenvalues by

$$\lambda(A) = \{\lambda_1, \ldots, \lambda_n\} \quad \text{and} \quad \lambda(\widetilde{A}) = \{\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_n\} \tag{3.1}$$

ordered so that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n, \quad \widetilde{\lambda}_1 \geq \widetilde{\lambda}_2 \geq \cdots \geq \widetilde{\lambda}_n. \tag{3.2}$$

THEOREM 3.1. *Let A and $\widetilde{A} = D^*AD$ be two $n \times n$ Hermitian matrices with eigenvalues* (3.1) *ordered as in* (3.2)*, where D is nonsingular. Then*

1. *there is a permutation τ of $\{1, 2, \ldots, n\}$ such that*

$$\sqrt{\sum_{i=1}^{n} \left[\varrho_2(\lambda_i, \widetilde{\lambda}_{\tau(i)})\right]^2} \leq \sqrt{\|I - \Sigma_{\mathrm{d}}\|_{\mathrm{F}}^2 + \|I - \Sigma_{\mathrm{d}}^{-1}\|_{\mathrm{F}}^2}, \tag{3.3}$$

where Σ_{d} is diagonal and its diagonal entries are D's singular values.

2. *if, in addition, A is nonnegative definite,*[1] *then*

$$\max_{1\leq i\leq n} \chi(\lambda_i, \widetilde{\lambda}_i) \leq \|D^* - D^{-1}\|_2, \tag{3.4}$$

$$\sqrt{\sum_{i=1}^{n} \left[\chi(\lambda_i, \widetilde{\lambda}_i)\right]^2} \leq \|D^* - D^{-1}\|_{\mathrm{F}}. \tag{3.5}$$

A proof of Theorem 3.1 will be given in section 5.

A corollary of (3.3) is

$$\sqrt{\sum_{i=1}^{n} \left[\varrho_2(\lambda_i, \widetilde{\lambda}_{\tau(i)})\right]^2} \leq \sqrt{\|I - D\|_{\mathrm{F}}^2 + \|I - D^{-1}\|_{\mathrm{F}}^2} \tag{3.3a}$$

by a well-known (absolute) perturbation theorem for singular values; see (4.7). On the other hand, (3.3a) leads to (3.3) as well by considering $U_{\mathrm{d}}^* A U_{\mathrm{d}}$ and $V_{\mathrm{d}}^* \widetilde{A} V_{\mathrm{d}} = \Sigma_{\mathrm{d}}(U_{\mathrm{d}}^* A U_{\mathrm{d}})\Sigma_{\mathrm{d}}$ instead, where

$$D = U_{\mathrm{d}} \Sigma_{\mathrm{d}} V_{\mathrm{d}}^* \tag{3.6}$$

is D's singular value decomposition (SVD) [15, p. 71]. It is also possible to relate the right-hand sides of (3.4) and (3.5) to the singular values of D, since for every unitarily invariant norm[2] $|||\cdot|||$,

$$|||D^* - D^{-1}||| = |||V_{\mathrm{d}}(\Sigma_{\mathrm{d}} - \Sigma_{\mathrm{d}}^{-1})U_{\mathrm{d}}^*||| = |||\Sigma_{\mathrm{d}} - \Sigma_{\mathrm{d}}^{-1}|||.$$

[1] Then $\widetilde{A}$ must be nonnegative definite as well.

[2] In this we follow Mirsky [30], Stewart and Sun [35], and Bhatia [3]. That a norm $|||\cdot|||$ is *unitarily invariant* on $\mathbb{C}^{m\times n}$ means that it also satisfies, besides the usual properties of any norm,

1. $|||UYV||| = |||Y|||$, for any $U \in \mathbb{U}_m$, and $V \in \mathbb{U}_n$;
2. $|||Y||| = \|Y\|_2$, for any $Y \in \mathbb{C}^{m\times n}$ with $\mathrm{rank}(Y) = 1$.

Two unitarily invariant norms most frequently used are the *spectral norm* $\|\cdot\|_2$ and the *Frobenius norm* $\|\cdot\|_{\mathrm{F}}$. Let $|||\cdot|||$ be a unitarily invariant norm on some matrix space. The following inequalities [35, p. 80] will be employed later in this paper:

$$|||WY||| \leq \|W\|_2 |||Y||| \quad \text{and} \quad |||YZ||| \leq |||Y|||\|Z\|_2.$$

The earliest relative perturbation result for eigenvalue problems goes back to a theorem due to Ostrowski [32] (see also [18, pp. 224–225]), though he did not interpret his theorem in the way we do now. Ostrowski proved that

for two $n\times n$ Hermitian matrices A and $\widetilde{A} = D^*AD$ with eigenvalues (3.1) *ordered as in* (3.2)*, where D is nonsingular, we have*

$$\sigma_{\min}(D)^2 \cdot \lambda_i \le \widetilde{\lambda}_i \le \sigma_{\max}(D)^2 \cdot \lambda_i \quad \textit{for } \ \le i \le n. \tag{3.7}$$

Inequalities (3.7) immediately imply a relative perturbation bound

$$\max_{1\le i\le n} \frac{|\widetilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \le \|I - D^*D\|_2.$$

This result of Ostrowski's is independent of (3.4). Both may be attainable for the scalar case ($n = 1$) or for the case when A and D are diagonal. Our bounds (3.3) and (3.5) are the first of their kind.

Roughly speaking, the classical perturbation theory for Hermitian matrix eigenvalue problems establishes one uniform bound for all differences $|\lambda_i - \widetilde{\lambda}_i|$ regardless of magnitudes of λ_i's. In this regard, we have the following.

Let both A and $\widetilde{A}$ be Hermitian. (*No special form of $\widetilde{A}$ is assumed.*) *Then for any unitarily invariant norm $|||\cdot|||$,*

$$|||\mathrm{diag}(\lambda_1 - \widetilde{\lambda}_1, \ldots, \lambda_n - \widetilde{\lambda}_n)||| \le |||A - \widetilde{A}|||. \tag{3.8}$$

There is a long history associated with this inequality; see Bhatia [3] for details. Theorem 3.1 extends (3.8) to the relative perturbation theory for $|||\cdot||| = \|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$. Two main differences between Theorem 3.1 and (3.8) are as follows.

1. Inequality (3.8) bounds the absolute differences $|\lambda_i - \widetilde{\lambda}_i|$. It is in fact the best possible as far as arbitrary perturbations are concerned. However, it may overestimate the differences $|\lambda_j - \widetilde{\lambda}_j|$ too much for eigenvalues λ_j of much smaller magnitudes than $\|A\|_2$ when perturbations have special structures such as multiplicative perturbations, for which it is possible that $|||A - \widetilde{A}|||$ is larger than $|\lambda_j - \widetilde{\lambda}_j|$ by many orders of magnitudes while, on the other hand, $D^*D \approx I$.
2. Theorem 3.1 exploits fully multiplicative perturbation structures by bounding directly the relative differences $\chi(\lambda_i, \widetilde{\lambda}_i)$ or $\varrho_2(\lambda_i, \widetilde{\lambda}_i)$ in terms of D's departures from unitary matrices $|||D^* - D^{-1}|||$ and $\sqrt{\|I - \Sigma_d\|_{\mathrm{F}}^2 + \|I - \Sigma_d^{-1}\|_{\mathrm{F}}^2}$. Thus, all eigenvalues of the same as or much smaller magnitudes than $\|A\|_2$ alike provably suffer small uncertainty as long as D's departures from unitary matrices are small.

Such arguments more or less apply to our other relative perturbation theorems in this paper in comparison to their counterparts in the classical absolute perturbation theory.

In Theorem 3.1, the perturbation to A is rather restrictive but *is* applicable to a more realistic situation when scaled A is much better conditioned. In Theorem 3.2, S is a scaling matrix, often highly graded and diagonal in practice, though the theorem does not assume this.

THEOREM 3.2. *Let $A = S^*HS$ and $\widetilde{A} = S^*\widetilde{H}S$ be two $n \times n$ nonnegative definite Hermitian matrices with eigenvalues* (3.1) *ordered as in* (3.2)*, and let $\Delta H = \widetilde{H} - H$.*

If $\|H^{-1}\|_2\|\Delta H\|_2 < 1$, *then*

$$\max_{1\le i\le n} \chi(\lambda_i, \widetilde{\lambda}_i) \le \|D - D^{-1}\|_2, \tag{3.9}$$

$$\le \frac{\|H^{-1}\|_2\|\Delta H\|_2}{\sqrt{1 - \|H^{-1}\|_2\|\Delta H\|_2}}, \tag{3.10}$$

$$\sqrt{\sum_{i=1}^{n} \Big[\chi(\lambda_i, \widetilde{\lambda}_i)\Big]^2} \le \|D - D^{-1}\|_{\mathrm{F}}, \tag{3.11}$$

$$\le \frac{\|H^{-1}\|_2\|\Delta H\|_{\mathrm{F}}}{\sqrt{1 - \|H^{-1}\|_2\|\Delta H\|_2}}, \tag{3.12}$$

where $D = \big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2}$.

Proof. Rewrite A and $\widetilde{A}$ as

$$\begin{aligned} A &= S^*HS = (H^{1/2}S)^*\, H^{1/2}S, \\ \widetilde{A} &= S^*H^{1/2}\big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)H^{1/2}S \\ &= \Big(\big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2}H^{1/2}S\Big)^* \big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2}H^{1/2}S. \end{aligned}$$

Set $B \stackrel{\text{def}}{=} H^{1/2}S$ and $\widetilde{B} \stackrel{\text{def}}{=} \big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2}H^{1/2}S$, then $A = B^*B$ and $\widetilde{A} = \widetilde{B}^*\widetilde{B}$. We have $\widetilde{B} = DB$, where $D = \big(I + H^{-1/2}(\Delta H)H^{-1/2}\big)^{1/2}$. Notice that

$$\lambda(A) = \lambda(B^*B) = \lambda(BB^*) \quad \text{and} \quad \lambda(\widetilde{A}) = \lambda(\widetilde{B}^*\widetilde{B}) = \lambda(\widetilde{B}\widetilde{B}^*),$$

and $\widetilde{B}\widetilde{B}^* = DBB^*D^*$. Applying Theorem 3.1 to BB^* and $\widetilde{B}\widetilde{B}^*$ yields both (3.9) and (3.11). Inequalities (3.10) and (3.12) follow from the fact that for any Hermitian matrix E with $\|E\|_2 < 1$ and for any unitarily invariant norm $\|\!|\cdot|\!\|$,

$$\|\!|(I+E)^{1/2} - (I+E)^{-1/2}|\!\| \le \|(I+E)^{-1/2}\|_2\|\!|E|\!\| \le \frac{\|\!|E|\!\|}{\sqrt{1 - \|E\|_2}}. \qquad \square$$

Inequality (3.10) can also be derived from the following bound essentially due to Demmel and Veselić [9] (see also Mathias [28]).

Let the conditions of Theorem 3.2 *hold. Then*

$$\max_{1\le i\le n} \frac{|\widetilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \le \|H^{-1}\|_2\|\Delta H\|_2. \tag{3.13}$$

To see how (3.13) leads to (3.10), we notice that[3]

$$\chi(\lambda_i, \widetilde{\lambda}_i) = \frac{|\widetilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \cdot \sqrt{\frac{\lambda_i}{\widetilde{\lambda}_i}} \le \frac{|\widetilde{\lambda}_i - \lambda_i|}{|\lambda_i|} \cdot \|D^{-1}\|_2$$

by Ostrowski's theorem (3.7) and that $\|D^{-1}\|_2 \le 1/\sqrt{1 - \|H^{-1}\|_2\|\Delta H\|_2}$.

Remark 3.1. Li [24] also considered extending Theorem 3.1 to diagonalizable matrices under multiplicative perturbations. But the bounds obtained in a recent paper [26] are better. Both Li [24] and Eisenstat and Ipsen [13] extended the classical Bauer–Fike theorem [2].

[3] $\lambda_i = 0$ if and only if $\widetilde{\lambda}_i = 0$, since A and $\widetilde{A}$ have the same number of zero eigenvalues, if any. So we only need to consider those i such that $\lambda_i \ne 0$.

4. Relative perturbation theorems for singular value problems. Throughout the section, $B, \widetilde{B} \in \mathbb{C}^{m\times n}$ and one is a perturbation of the other. (We shall assume, without loss of generality, that $m \ge n$ in this section.) Denote their singular values by

$$\sigma(B) = \{\sigma_1, \ldots, \sigma_n\} \quad \text{and} \quad \sigma(\widetilde{B}) = \{\widetilde{\sigma}_1, \ldots, \widetilde{\sigma}_n\} \tag{4.1}$$

ordered so that

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n \ge 0, \qquad \widetilde{\sigma}_1 \ge \widetilde{\sigma}_2 \ge \cdots \ge \widetilde{\sigma}_n \ge 0. \tag{4.2}$$

THEOREM 4.1. *Let B and $\widetilde{B} = D_1^* B D_2$ be two $m \times n$ matrices with singular values* (4.1) *ordered as in* (4.2)*, where D_1 and D_2 are square and nonsingular. If $\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2 < 32$, then*

$$\max_{1\le i\le n} \chi(\sigma_i, \widetilde{\sigma}_i) \le \frac{1}{2} \cdot \frac{\|D_1^* - D_1^{-1}\|_2 + \|D_2^* - D_2^{-1}\|_2}{1 - \frac{1}{32}\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2}, \tag{4.3}$$

$$\sqrt{\sum_{i=1}^{n} [\chi(\sigma_i, \widetilde{\sigma}_i)]^2} \le \frac{1}{2} \cdot \frac{\|D_1^* - D_1^{-1}\|_{\mathrm{F}} + \|D_2^* - D_2^{-1}\|_{\mathrm{F}}}{1 - \frac{1}{32}\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2}. \tag{4.4}$$

A proof of Theorem 4.1 will be given in section 6.

The restriction $\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2 < 32$, though mild, is unpleasant. But we argue that neither this restriction nor the factor $\left(1 - \frac{1}{32}\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2\right)^{-1}$ plays any visible role for any applications where one might expect that perturbing B to $\widetilde{B} = D_1^* B D_2$ retains any significant digits of B's singular values. Our arguments go as follows.

1. For the ease of explanation, consider the case when B and D_j are diagonal. In order for each of B's singular values to have at least one significant decimal digit the same as that of the corresponding $\widetilde{B}$'s, it is necessary that[4]

$$0.9 \le \sigma_{\min}(D_j) \le \sigma_{\max}(D_j) \le 1.05 \tag{4.5}$$

which imply that $\|D_j^* - D_j^{-1}\|_2 \le 0.2$, and thus the factor

$$\left(1 - \frac{1}{32}\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2\right)^{-1} \le 1.01.$$

2. In fact, the restriction $\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2 < 32$ is satisfied and the factor is almost 1 even for D_j's singular values being fairly away from 1. It can be seen that

$$\|D_j^* - D_j^{-1}\|_2 \le 1 \text{ if } 0.618 \approx \frac{\sqrt{5}-1}{2} \le \sigma_{\min}(D_j) \le \sigma_{\max}(D_j) \le \frac{\sqrt{5}+1}{2} \approx 1.618,$$

under which circumstances the unpleasant factor is

$$\left(1 - \frac{1}{32}\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2\right)^{-1} \le 32/31 \approx 1.03.$$

[4]This is for the worse case in the sense that if (4.5) is violated, then there are D_j's such that some of the B's singular values retain no significant decimal digits at all under the perturbations.

3. In applications where $\|D_j^* - D_j^{-1}\|_2 \ll 1$, the quantity $\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2$ is of second order. Then the restriction and the factor act as if they were not there. Even more in some applications, as in Corollary 4.2, one of the D_j's is I for which the restriction and the factor disappear completely.

Eisenstat and Ipsen [12] obtained the following result which is essentially a consequence of Ostrowski's theorem (see inequalities (3.7)) and which can also be seen from known inequalities for singular values of a product of two matrices:[5]

Let the conditions of Theorem 4.1*, except* $\|D_1^* - D_1^{-1}\|_2 \|D_2^* - D_2^{-1}\|_2 < 32$*, hold. We have*

$$\sigma_{\min}(D_1)\sigma_{\min}(D_2) \cdot \sigma_i \le \widetilde{\sigma}_i \le \sigma_{\max}(D_1)\sigma_{\max}(D_2) \cdot \sigma_i \quad \text{for } 1 \le i \le n. \tag{4.6}$$

Inequalities (4.6) imply immediately the following relative perturbation bound:

$$\max_{1 \le i \le n} \frac{|\widetilde{\sigma}_i - \sigma_i|}{\sigma_i} \le \max\{|1 - \sigma_{\min}(D_1)\sigma_{\min}(D_2)|,\ |1 - \sigma_{\max}(D_1)\sigma_{\max}(D_2)|\}.$$

The classical perturbation theory for singular value problems establishes one uniform bound for all differences $\sigma_i - \widetilde{\sigma}_i$, regardless of magnitudes of σ_i's. The following theorem was established by Mirsky [30], based on results from Lidskii [27] and Wielandt [39].

For any unitarily invariant norm $|||\cdot|||$*, we have*

$$|||\mathrm{diag}(\sigma_1 - \widetilde{\sigma}_1, \ldots, \sigma_n - \widetilde{\sigma}_n)||| \le |||B - \widetilde{B}|||. \tag{4.7}$$

(*No special form of* $\widetilde{B}$ *is assumed.*)

A possible application of Theorem 4.1 is related to *deflation* in computing SVD of a bidiagonal matrix. For more details, the reader is referred to [6, 8, 12, 29].

COROLLARY 4.2. *Assume, in Theorem* 4.1*, that one of* D_1 *and* D_2 *is the identity matrix and the other takes the form*

$$D = \begin{pmatrix} I & X \\ & I \end{pmatrix},$$

where X *is a matrix of suitable dimensions. Then*

$$\max_{1 \le i \le n} \chi(\sigma_i, \widetilde{\sigma}_i) \le \frac{1}{2}\|X\|_2, \tag{4.8}$$

$$\sqrt{\sum_{i=1}^{n} [\chi(\sigma_i, \widetilde{\sigma}_i)]^2} \le \frac{1}{\sqrt{2}}\|X\|_{\mathrm{F}}. \tag{4.9}$$

Proof. Notice that

$$D^* - D^{-1} = \begin{pmatrix} I & \\ X^* & I \end{pmatrix} - \begin{pmatrix} I & -X \\ & I \end{pmatrix} = \begin{pmatrix} & X \\ X^* & \end{pmatrix},$$

[5] Arranging the singular values of a matrix in the decreasing order, we have (see, e.g., [19])

(the ith singular value of XY) $\le$ (the ith singular value of X) $\cdot \|Y\|_2$.

and thus $\|D^* - D^{-1}\|_2 = \|X\|_2$ and $\|D^* - D^{-1}\|_{\rm F} = \sqrt{2}\|X\|_{\rm F}$. ∎

Eisenstat and Ipsen [12] showed that

$$|\widetilde{\sigma}_i - \sigma_i| \le \|X\|_2\sigma_i, \quad \text{or equivalently} \quad \left|\frac{\widetilde{\sigma}_i}{\sigma_i} - 1\right| \le \|X\|_2. \tag{4.10}$$

Our inequality (4.8) is sharper by roughly a factor of 1/2, as long as $\|X\|_2$ is small. As a matter of fact, it follows from (4.8) and Proposition 2.1 that if $\|X\|_2 < 4$, then

$$\left|\frac{\widetilde{\sigma}_i}{\sigma_i} - 1\right| \le \left(\frac{\|X\|_2}{4} + \sqrt{1 + \frac{\|X\|_2^2}{16}}\right)\frac{\|X\|_2}{2} = \frac{\|X\|_2}{2} + O\left(\left(\frac{\|X\|_2}{4}\right)^2\right).$$

Our inequality (4.9) is the first of its kind.

THEOREM 4.3. *Let B and $\widetilde{B} = D_1^*BD_2$ be two $m \times n$ matrices with singular values* (4.1) *ordered as in* (4.2)*, where D_1 and D_2 are square and nonsingular. Then*

$$\max_{1\le i\le n} \varrho_p(\sigma_i, \widetilde{\sigma}_i) \le \frac{1}{2^{1+1/p}}\left(\|D_1^* - D_1^{-1}\|_2 + \|D_2^* - D_2^{-1}\|_2\right), \tag{4.11}$$

$$\sqrt{\sum_{i=1}^n [\varrho_p(\sigma_i, \widetilde{\sigma}_i)]^2} \le \frac{1}{2^{1+1/p}}\left(\|D_1^* - D_1^{-1}\|_{\rm F} + \|D_2^* - D_2^{-1}\|_{\rm F}\right). \tag{4.12}$$

A straightforward combination of Proposition 2.2 and Theorem 4.1 will lead to bounds that are slightly weaker than those in Theorem 4.3 by a factor of

$$\left(1 - \frac{1}{32}\|D_1^* - D_1^{-1}\|_2\|D_2^* - D_2^{-1}\|_2\right)^{-1}.$$

A proof of Theorem 4.3 will be given in section 6.

Again we shall now consider a more realistic situation when scaled B is much better conditioned. In Theorem 4.4 below, S is a scaling matrix, often highly graded and diagonal in practice, though the theorem does not assume this.

THEOREM 4.4. *Let $B = GS$ and $\widetilde{B} = \widetilde{G}S$ be two $n \times n$ matrices with singular values* (4.1) *ordered as in* (4.2)*, where G and $\widetilde{G}$ are nonsingular, and let $\Delta G = \widetilde{G} - G$. If $\|\Delta G\|_2\|G^{-1}\|_2 < 1$, then*

$$\max_{1\le i\le n} \chi(\sigma_i, \widetilde{\sigma}_i) \le \frac{1}{2}\left\|\left(I + (\Delta G)G^{-1}\right)^* - \left(I + (\Delta G)G^{-1}\right)^{-1}\right\|_2, \tag{4.13}$$

$$\le \left(1 + \frac{1}{1 - \|G^{-1}\|_2\|\Delta G\|_2}\right)\frac{\|G^{-1}\|_2\|\Delta G\|_2}{2}, \tag{4.14}$$

$$\sqrt{\sum_{i=1}^n [\chi(\sigma_i, \widetilde{\sigma}_i)]^2} \le \frac{1}{2}\left\|\left(I + (\Delta G)G^{-1}\right)^* - \left(I + (\Delta G)G^{-1}\right)^{-1}\right\|_{\rm F}, \tag{4.15}$$

$$\le \left(1 + \frac{1}{1 - \|G^{-1}\|_2\|\Delta G\|_2}\right)\frac{\|G^{-1}\|_2\|\Delta G\|_{\rm F}}{2}. \tag{4.16}$$

Proof. Write

$$\widetilde{B} = (G + \Delta G)S = (I + (\Delta G)G^{-1})GS = DB, \tag{4.17}$$

where $D = I + (\Delta G)G^{-1}$. Now, applying Theorem 4.1 to B and $\widetilde{B} = DB$ yields both (4.13) and (4.15). We notice that

$$(I+E)^* - (I+E)^{-1} = I + E^* - \sum_{i=0}^{\infty}(-1)^i E^i = E^* + E + E\sum_{i=2}^{\infty}(-1)^i E^{i-1},$$

where $E = (\Delta G)G^{-1}$ and $\|E\|_2 \le \|G^{-1}\|_2\|\Delta G\|_2 < 1$; therefore, for any unitarily invariant norm $|||\cdot|||$,

$$\begin{aligned}
|||(I+E)^* - (I+E)^{-1}||| &\le |||E+E^*||| + |||E|||\sum_{i=1}^{\infty}\|E\|_2^i \\
&= \left(\frac{|||E+E^*|||}{|||E|||} + \frac{\|E\|_2}{1-\|E\|_2}\right)|||E||| \qquad (4.18)\\
&\le \left(1 + \frac{1}{1-\|E\|_2}\right)|||E|||. \qquad (4.19)
\end{aligned}$$

An application of (4.19) for $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{F}}$ completes the proof. ∎

Equation (4.17) also makes (4.6) applicable and leads to the following.

Let the conditions of Theorem 4.4 *hold. We have*

$$\max_{1\le i\le n}\frac{|\widetilde{\sigma}_i - \sigma_i|}{\sigma_i} \le \|G^{-1}\|_2\|\Delta G\|_2. \qquad (4.20)$$

This inequality also follows from [10, Theorem 1.1]. Inequality (4.14) can actually be derived from (4.20) as follows. Notice that

$$\chi(\sigma_i,\widetilde{\sigma}_i) = \frac{|\widetilde{\sigma}_i - \sigma_i|}{|\sigma_i|}\cdot\sqrt{\frac{\sigma_i}{\widetilde{\sigma}_i}} \le \frac{|\widetilde{\sigma}_i - \sigma_i|}{|\sigma_i|}\cdot\|D^{-1}\|_2^{1/2},$$

and that

$$\|D^{-1}\|_2^{1/2} \le \frac{1}{\sqrt{1-\|G^{-1}\|_2\|\Delta G\|_2}} \le \frac{1}{2}\left(1 + \frac{1}{1-\|G^{-1}\|_2\|\Delta G\|_2}\right).$$

Remark 4.1. When $(\Delta G)G^{-1}$ is nearly skew Hermitian, (4.13) and (4.15) lead to bounds that are much better than (4.14) and (4.16). This can be seen from (4.18):

Under the conditions of Theorem 4.4*, we have*

$$\max_{1\le i\le n}\chi(\sigma_i,\widetilde{\sigma}_i) \le \left(\frac{\|(\Delta G)G^{-1} + G^{-*}(\Delta G)^*\|_2}{\|(\Delta G)G^{-1}\|_2} + \frac{\|(\Delta G)G^{-1}\|_2}{1-\|(\Delta G)G^{-1}\|_2}\right)\frac{\|(\Delta G)G^{-1}\|_2}{2},$$

$$\sqrt{\sum_{i=1}^{n}[\chi(\sigma_i,\widetilde{\sigma}_i)]^2} \le \left(\frac{\|(\Delta G)G^{-1} + G^{-*}(\Delta G)^*\|_{\mathrm{F}}}{\|(\Delta G)G^{-1}\|_{\mathrm{F}}} + \frac{\|(\Delta G)G^{-1}\|_2}{1-\|(\Delta G)G^{-1}\|_2}\right)\frac{\|(\Delta G)G^{-1}\|_{\mathrm{F}}}{2}.$$

Now if $(\Delta G)G^{-1}$ is nearly skew Hermitian, then $\chi(\sigma_i,\widetilde{\sigma}_i) = o\left(\|(\Delta G)G^{-1}\|_2\right)$; moreover,

$$\|(\Delta G)G^{-1} + G^{-*}(\Delta G)^*\|_2 = O\left(\|(\Delta G)G^{-1}\|_2^2\right) \Rightarrow \chi(\sigma_i,\widetilde{\sigma}_i) = O\left(\|(\Delta G)G^{-1}\|_2^2\right).$$

Remark 4.2. Theorem 4.4 can be extended to nonsquare matrices. Assume $B = GS$ and $\widetilde{B} = \widetilde{G}S$ are $m \times n$ ($m \ge n$); S is a scaling matrix and both G and $\widetilde{G}$ are $m \times n$; G has full column rank. Let $G^\dagger = (G^*G)^{-1}G^*$ be the pseudo-inverse of G. Notice that $G^\dagger G = I$. We have

$$\widetilde{B} = \widetilde{G}S = (G + \Delta G)S = (I + (\Delta G)G^\dagger)GS = (I + (\Delta G)G^\dagger)B \equiv DB.$$

Now, apply Theorem 4.1 to B and $\widetilde{B} = DB$.

5. Proof of Theorem 3.1. We need a little preparation first. A matrix $Z = (z_{ij}) \in \mathbb{R}^{n\times n}$ is *doubly stochastic* if all $z_{ij} \ge 0$ and

$$\sum_{k=1}^{n} z_{ik} = \sum_{k=1}^{n} z_{kj} = 1 \quad \text{for } i, j = 1, 2, \ldots, n.$$

Using a Birkhoff theorem [4] (see also [18, pp. 527–528]) and the technique of Hoffman and Wielandt [17] (see also [35, p. 190]), we can prove the following.

LEMMA 5.1. *Let $Z = (z_{ij})$ be an $n \times n$ doubly stochastic matrix, and let $M = (m_{ij}) \in \mathbb{C}^{n\times n}$. Then there exists a permutation τ of $\{1, 2, \ldots, n\}$ such that*

$$\sum_{i,\,j=1}^{n} |m_{ij}| z_{ij} \ge \sum_{i=1}^{n} |m_{i\tau(i)}|.$$

For $X \in \mathbb{C}^{m\times n}$, we introduce the following notation for a $k \times \ell$ submatrix of $X = (x_{ij})$:

$$X\begin{pmatrix} i_1 \ldots i_k \\ j_1 \ldots j_\ell \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} x_{i_1 j_1} & x_{i_1 j_2} & \cdots & x_{i_1 j_\ell} \\ x_{i_2 j_1} & x_{i_2 j_2} & \cdots & x_{i_2 j_\ell} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i_k j_1} & x_{i_k j_2} & \cdots & x_{i_k j_\ell} \end{pmatrix}, \tag{5.1}$$

where $1 \le i_1 < \cdots < i_k \le n$ and $1 \le j_1 < \cdots < j_\ell \le n$. The following lemma is due to Li [22, pp. 207–208]

LEMMA 5.2 (see Li [22]). *Suppose that $X \in \mathbb{C}^{n\times n}$ is nonsingular, and $1 \le i_1 < \cdots < i_k \le n$ and $1 \le j_1 < \cdots < j_\ell \le n$, and $k + \ell > n$. Then*

$$\left\| X\begin{pmatrix} i_1 \cdots i_k \\ j_1 \cdots j_\ell \end{pmatrix} \right\|_2 \ge \|X^{-1}\|_2^{-1}.$$

Moreover, if X is unitary, then

$$\left\| X\begin{pmatrix} i_1 \cdots i_k \\ j_1 \cdots j_\ell \end{pmatrix} \right\|_2 = 1.$$

Proof of Theorem 3.1. We shall prove (3.3) first. Due to the argument we made right after Theorem 3.1, it suffices for us to prove (3.3a). Let the eigen decompositions of A and $\widetilde{A}$ be

$$A = U\Lambda U^* \quad \text{and} \quad \widetilde{A} = \widetilde{U}\widetilde{\Lambda}\widetilde{U}^*,$$

where U and $\widetilde{U}$ are unitary and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $\widetilde{\Lambda} = \mathrm{diag}(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_n)$. Notice that

$$A - \widetilde{A} = A - D^*AD = A - AD + AD - D^*AD = A(I - D) + (D^{-*} - I)\widetilde{A}.$$

Pre- and postmultiply the equations by U^* and $\widetilde{U}$, respectively, to get

$$\Lambda U^*\widetilde{U} - U^*\widetilde{U}\widetilde{\Lambda} = \Lambda U^*(I - D)\widetilde{U} + U^*(D^{-*} - I)\widetilde{U}\widetilde{\Lambda}. \tag{5.2}$$

Set

$$Q \stackrel{\mathrm{def}}{=} U^*\widetilde{U} = (q_{ij}), \quad E \stackrel{\mathrm{def}}{=} U^*(I - D)\widetilde{U} = (e_{ij}), \quad \widetilde{E} \stackrel{\mathrm{def}}{=} U^*(D^{-*} - I)\widetilde{U} = (\widetilde{e}_{ij}).$$

Then (5.2) reads $\Lambda Q - Q\widetilde{\Lambda} = \Lambda E + \widetilde{E}\widetilde{\Lambda}$, or componentwise $\lambda_i q_{ij} - q_{ij}\widetilde{\lambda}_j = \lambda_i e_{ij} + \widetilde{e}_{ij}\widetilde{\lambda}_j$, so

$$|(\lambda_i - \widetilde{\lambda}_j)q_{ij}|^2 = |\lambda_i e_{ij} + \widetilde{e}_{ij}\widetilde{\lambda}_j|^2 \le (|\lambda_i|^2 + |\widetilde{\lambda}_j|^2)(|e_{ij}|^2 + |\widetilde{e}_{ij}|^2),$$

which yields[6] $[\varrho_2(\lambda_i, \widetilde{\lambda}_j)]^2|q_{ij}|^2 \le |e_{ij}|^2 + |\widetilde{e}_{ij}|^2$. Hence

$$\begin{aligned}\sum_{i,\,j=1}^{n} \Big[\varrho_2(\lambda_i, \widetilde{\lambda}_j)\Big]^2 |q_{ij}|^2 &\le \|U^*(I - D)\widetilde{U}\|_{\mathrm{F}}^2 + \|U^*(D^{-*} - I)\widetilde{U}\|_{\mathrm{F}}^2 \\ &= \|I - D\|_{\mathrm{F}}^2 + \|D^{-*} - I\|_{\mathrm{F}}^2.\end{aligned}$$

The matrix $(|q_{ij}|^2)_{n\times n}$ is a doubly stochastic matrix. The above inequality and Lemma 5.1 imply that

$$\sum_{i=1}^{n} \Big[\varrho_2(\lambda_i, \widetilde{\lambda}_{\tau(i)})\Big]^2 \le \|I - D\|_{\mathrm{F}}^2 + \|D^{-*} - I\|_{\mathrm{F}}^2$$

for some permutation τ of $\{1, 2, \ldots, n\}$. This is (3.3a).

We now prove (3.4) and (3.5). Suppose that A is nonnegative definite. There is a matrix $B \in \mathbb{C}^{n\times n}$ such that $A = B^*B$. With this B, $\widetilde{A} = D^*AD = D^*B^*BD = \widetilde{B}^*\widetilde{B}$, where $\widetilde{B} = BD$. Let SVDs of B and $\widetilde{B}$ be

$$B = U\Lambda^{1/2}V^* \quad \text{and} \quad \widetilde{B} = \widetilde{U}\widetilde{\Lambda}^{1/2}\widetilde{V}^*,$$

where $\Lambda^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n})$ and $\widetilde{\Lambda}^{1/2} = \mathrm{diag}\Big(\sqrt{\widetilde{\lambda}_1}, \sqrt{\widetilde{\lambda}_2}, \ldots, \sqrt{\widetilde{\lambda}_n}\Big)$. In what follows, we actually work with BB^* and $\widetilde{B}\widetilde{B}^*$, rather than $A = B^*B$ and $\widetilde{A} = \widetilde{B}^*\widetilde{B}$ themselves. We have

$$\widetilde{B}\widetilde{B}^* - BB^* = \widetilde{B}D^*B^* - \widetilde{B}D^{-1}B^* = \widetilde{B}(D^* - D^{-1})B^*.$$

Pre- and postmultiply the above equations by $\widetilde{U}^*$ and U, respectively, to get

$$\widetilde{\Lambda}\widetilde{U}^*U - \widetilde{U}^*U\Lambda = \widetilde{\Lambda}^{1/2}\widetilde{V}^*(D^* - D^{-1})V\Lambda^{1/2}. \tag{5.3}$$

Write $Q \stackrel{\mathrm{def}}{=} \widetilde{U}^*U = (q_{ij})$. Equation (5.3) implies

$$\|D^* - D^{-1}\|_{\mathrm{F}}^2 = \|\widetilde{V}^*(D^* - D^{-1})V\|_{\mathrm{F}}^2 = \sum_{i,\,j=1}^{n} \frac{|\widetilde{\lambda}_i - \lambda_j|}{\sqrt{\widetilde{\lambda}_i\lambda_j}}|q_{ij}|^2.$$

[6] This inequality still holds even if $\lambda_i = \widetilde{\lambda}_j = 0$ because of our convention $0/0 = 0$; see section 2.

Since $(|q_{ij}|^2)_{n\times n}$ is a doubly stochastic matrix, an application of Lemma 5.1 and Proposition 2.4 concludes the proof of (3.5). To confirm (3.4), let k be the index such that

$$\eta \stackrel{\rm def}{=} \max_{1\le i\le n} \chi(\lambda_i, \widetilde{\lambda}_i) = \chi(\lambda_k, \widetilde{\lambda}_k).$$

If $\eta = 0$, no proof is necessary. Assume $\eta > 0$. Also assume, without loss of generality, that

$$\lambda_k > \widetilde{\lambda}_k \ge 0.$$

Partition U, V, $\widetilde{U}$, $\widetilde{V}$ as follows:

$$U = \begin{array}{c} \overset{k}{\big(} U_1 \end{array} \begin{array}{c} \overset{n-k}{\quad} U_2 \big) \end{array}, V = \begin{array}{c} \overset{k}{\big(} V_1 \end{array} \begin{array}{c} \overset{n-k}{\quad} V_2 \big) \end{array}, \widetilde{U} = \begin{array}{c} \overset{k-1}{\big(} \widetilde{U}_1 \end{array} \begin{array}{c} \overset{n-k+1}{\quad} \widetilde{U}_2 \big) \end{array}, \widetilde{V} = \begin{array}{c} \overset{k-1}{\big(} \widetilde{V}_1 \end{array} \begin{array}{c} \overset{n-k+1}{\quad} \widetilde{V}_2 \big) \end{array},$$

and write $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$ and $\widetilde{\Lambda} = \mathrm{diag}(\widetilde{\Lambda}_1, \widetilde{\Lambda}_2)$, where $\Lambda_1 \in \mathbb{R}^{k\times k}$ and $\widetilde{\Lambda}_1 \in \mathbb{R}^{(k-1)\times(k-1)}$. It follows from (5.3) that

$$\widetilde{\Lambda}_2\widetilde{U}_2^*U_1 - \widetilde{U}_2^*U_1\Lambda_1 = \widetilde{\Lambda}_2^{1/2}\widetilde{V}_2^*(D^* - D^{-1})V_1\Lambda_1^{1/2}.$$

Postmultiply this equation by Λ_1^{-1} to get

$$\widetilde{\Lambda}_2\widetilde{U}_2^*U_1\Lambda_1^{-1} - \widetilde{U}_2^*U_1 = \widetilde{\Lambda}_2^{1/2}\widetilde{V}_2^*(D^* - D^{-1})V_1\Lambda_1^{-1/2}. \tag{5.4}$$

Lemma 5.2 implies that $\|\widetilde{U}_2^*U_1\|_2 = 1$ since $\widetilde{U}_2^*U_1$ is an $(n-k+1)\times k$ submatrix of unitary $\widetilde{U}^*U$ and $k + (n-k+1) = n+1 > n$. Bearing in mind that $\|\widetilde{\Lambda}_2\|_2 = \widetilde{\lambda}_k = \|\widetilde{\Lambda}_2^{1/2}\|_2^2$ and $\|\Lambda_1^{-1}\|_2 = 1/\lambda_k = \|\Lambda_1^{-1/2}\|_2^2$, we have

$$\begin{aligned}
1 - \frac{\widetilde{\lambda}_k}{\lambda_k} &= \left\|\widetilde{U}_2^*U_1\right\|_2 - \|\widetilde{\Lambda}_2\|_2 \left\|\widetilde{U}_2^*U_1\right\|_2 \|\Lambda_1^{-1}\|_2 \\
&\le \left\|\widetilde{U}_2^*U_1\right\|_2 - \left\|\widetilde{\Lambda}_2\widetilde{U}_2^*U_1\Lambda_1^{-1}\right\|_2 \\
&\le \left\|\widetilde{U}_2^*U_1 - \widetilde{\Lambda}_2\widetilde{U}_2^*U_1\Lambda_1^{-1}\right\|_2 \\
&= \left\|\widetilde{\Lambda}_2^{1/2}\widetilde{V}_2^*(D^* - D^{-1})V_1\Lambda_1^{-1/2}\right\|_2 \quad \text{(by (5.4))} \\
&\le \|\widetilde{\Lambda}_2^{1/2}\|_2 \left\|\widetilde{V}_2^*(D^* - D^{-1})V_1\right\|_2 \|\Lambda_1^{-1/2}\|_2 \\
&= \sqrt{\frac{\widetilde{\lambda}_k}{\lambda_k}} \left\|\widetilde{V}_2^*(D^* - D^{-1})V_1\right\|_2 \\
&\le \sqrt{\frac{\widetilde{\lambda}_k}{\lambda_k}} \|D^* - D^{-1}\|_2,
\end{aligned}$$

an immediate consequence of which is (3.4). ∎

6. Proofs of Theorems 4.1 and 4.3. We need the following lemma regarding the relative distance χ.

LEMMA 6.1.

1. *If* $0 \le \alpha \le \beta \le \widetilde{\beta} \le \widetilde{\alpha}$, *then* $\chi(\alpha, \widetilde{\alpha}) \ge \chi(\beta, \widetilde{\beta})$.

2. *If* $\alpha,\ \widetilde{\alpha} \geq 0$, *then* $2\,\chi(\alpha,\widetilde{\alpha}) \leq \chi(\alpha^2,\widetilde{\alpha}^2)$.
3. *For* $\alpha,\ \beta,\ \gamma \geq 0$, *we have*

$$\chi(\alpha,\gamma) \leq \chi(\alpha,\beta) + \chi(\beta,\gamma) + \frac{1}{8}\chi(\alpha,\beta)\chi(\beta,\gamma)\chi(\alpha,\gamma). \tag{6.1}$$

Thus if $\chi(\alpha,\beta)\chi(\beta,\gamma) < 8$ *also, then*

$$\chi(\alpha,\gamma) \leq \frac{\chi(\alpha,\beta)+\chi(\beta,\gamma)}{1-\frac{1}{8}\chi(\alpha,\beta)\chi(\beta,\gamma)}.$$

Proof. To prove the first inequality, we notice that function $\frac{1}{x}-x$ is monotonically decreasing for $0 \leq x \leq 1$, and that $0 \leq \alpha/\widetilde{\alpha} \leq \beta/\widetilde{\beta} \leq 1$. Thus

$$\chi(\alpha,\widetilde{\alpha}) = \frac{1}{\sqrt{\alpha/\widetilde{\alpha}}} - \sqrt{\alpha/\widetilde{\alpha}} \geq \frac{1}{\sqrt{\beta/\widetilde{\beta}}} - \sqrt{\beta/\widetilde{\beta}} = \chi(\beta,\widetilde{\beta}),$$

as was to be shown. If $\alpha,\ \widetilde{\alpha} \geq 0$, then

$$\chi(\alpha^2,\widetilde{\alpha}^2) = \chi(\alpha,\widetilde{\alpha})\frac{|\alpha+\widetilde{\alpha}|}{\sqrt{|\alpha\widetilde{\alpha}|}} = \chi(\alpha,\widetilde{\alpha})\frac{\alpha+\widetilde{\alpha}}{\sqrt{\alpha\widetilde{\alpha}}} \geq \chi(\alpha,\widetilde{\alpha})\frac{2\sqrt{\alpha\widetilde{\alpha}}}{\sqrt{\alpha\widetilde{\alpha}}} = 2\chi(\alpha,\widetilde{\alpha}),$$

which confirms the second inequality.

For the third inequality (6.1), without loss of generality, we may assume $0 \leq \alpha \leq \gamma$. Now if $\beta \leq \alpha$ or $\gamma \leq \beta$, we have by the first inequality

$$\chi(\alpha,\gamma) \leq \begin{cases} \chi(\beta,\gamma) \leq \chi(\alpha,\beta)+\chi(\beta,\gamma), & \text{if } \beta \leq \alpha, \\ \chi(\alpha,\beta) \leq \chi(\alpha,\beta)+\chi(\beta,\gamma), & \text{if } \gamma \leq \beta, \end{cases}$$

so (6.1) holds. Consider the case $0 \leq \alpha \leq \beta \leq \gamma$. It can be verified that

$$\chi(\alpha,\gamma) = \chi(\alpha,\beta)+\chi(\beta,\gamma)+\chi(\sqrt{\alpha},\sqrt{\beta})\chi(\sqrt{\beta},\sqrt{\gamma})\chi(\sqrt{\alpha},\sqrt{\gamma}).$$

Inequality (6.1) follows by applying the second inequality. ∎

Proofs of Theorems 4.1 *and* 4.3. Set $\widehat{B} = BD_2$ and denote its singular values by $\widehat{\sigma}_1 \geq \widehat{\sigma}_2 \geq \cdots \geq \widehat{\sigma}_n$. Apply Theorem 3.1 to B^*B and $\widehat{B}^*\widehat{B} = D_2^*B^*BD_2$ to get

$$\max_{1\leq i\leq n}\chi(\sigma_i^2,\widehat{\sigma}_i^2) \leq \|D_2^* - D_2^{-1}\|_2 \quad \text{and} \quad \sqrt{\sum_{i=1}^n\left[\chi(\sigma_i^2,\widehat{\sigma}_i^2)\right]^2} \leq \|D_2^* - D_2^{-1}\|_{\mathrm{F}}.$$

Now apply the second inequality of Lemma 6.1 to obtain

$$\max_{1\leq i\leq n}\chi(\sigma_i,\widehat{\sigma}_i) \leq \frac{1}{2}\|D_2^* - D_2^{-1}\|_2 \quad \text{and} \quad \sqrt{\sum_{i=1}^n\left[\chi(\sigma_i,\widehat{\sigma}_i)\right]^2} \leq \frac{1}{2}\|D_2^* - D_2^{-1}\|_{\mathrm{F}}. \tag{6.2}$$

Similarly for $\widehat{B} = BD_2$ and $\widetilde{B} = D_1^*BD_2 = D_1^*\widehat{B}$, we have

$$\max_{1\leq i\leq n}\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i) \leq \frac{1}{2}\|D_1^* - D_1^{-1}\|_2 \quad \text{and} \quad \sqrt{\sum_{i=1}^n\left[\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i)\right]^2} \leq \frac{1}{2}\|D_1^* - D_1^{-1}\|_{\mathrm{F}}. \tag{6.3}$$

The first inequalities in (6.2) and (6.3), and the assumptions of Theorem 4.1, imply

$$\chi(\sigma_i,\widehat{\sigma}_i)\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i) \le \frac{1}{4}\|D_1^*-D_1^{-1}\|_2\|D_2^*-D_2^{-1}\|_2 < \frac{1}{4}\times 32 = 8.$$

By Lemma 6.1, we have

$$\begin{aligned}
\chi(\sigma_i,\widetilde{\sigma}_i) &\le \frac{\chi(\sigma_i,\widehat{\sigma}_i)+\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i)}{1-\frac{1}{8}\chi(\sigma_i,\widehat{\sigma}_i)\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i)}\\
&\le \frac{1}{2}\cdot\frac{\|D_1^*-D_1^{-1}\|_2+\|D_2^*-D_2^{-1}\|_2}{1-\frac{1}{32}\|D_1^*-D_1^{-1}\|_2\|D_2^*-D_2^{-1}\|_2},\\
\sqrt{\sum_{i=1}^n[\chi(\sigma_i,\widetilde{\sigma}_i)]^2} &\le \sqrt{\sum_{i=1}^n\left[\frac{\chi(\sigma_i,\widehat{\sigma}_i)+\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i)}{1-\frac{1}{8}\chi(\sigma_i,\widehat{\sigma}_i)\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i)}\right]^2}\\
&\le \frac{\sqrt{\sum_{i=1}^n[\chi(\sigma_i,\widehat{\sigma}_i)]^2}+\sqrt{\sum_{i=1}^n[\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i)]^2}}{1-\frac{1}{8}\max\limits_{1\le i\le n}\chi(\sigma_i,\widehat{\sigma}_i)\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i)}\\
&\le \frac{1}{2}\cdot\frac{\|D_1^*-D_1^{-1}\|_{\rm F}+\|D_2^*-D_2^{-1}\|_{\rm F}}{1-\frac{1}{32}\|D_1^*-D_1^{-1}\|_2\|D_2^*-D_2^{-1}\|_2},
\end{aligned}$$

as expected. This completes the proof of Theorem 4.1. To prove Theorem 4.3, we notice that

$$\begin{aligned}
\varrho_p(\sigma_i,\widetilde{\sigma}_i) &\le \varrho_p(\sigma_i,\widehat{\sigma}_i)+\varrho_p(\widehat{\sigma}_i,\widetilde{\sigma}_i) && (\varrho_p \text{ is a metric on } \mathbb{R})\\
&\le 2^{-1/p}\chi(\sigma_i,\widehat{\sigma}_i)+2^{-1/p}\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i) && (\text{by Proposition 2.2})\\
&\le 2^{-1-1/p}\left(\|D_2^*-D_2^{-1}\|_2+\|D_1^*-D_1^{-1}\|_2\right) && (\text{by (6.2) and (6.3)})
\end{aligned}$$

and

$$\begin{aligned}
\sqrt{\sum_{i=1}^n[\varrho_p(\sigma_i,\widetilde{\sigma}_i)]^2} &\le \sqrt{\sum_{i=1}^n[\varrho_p(\sigma_i,\widehat{\sigma}_i)+\varrho_p(\widehat{\sigma}_i,\widetilde{\sigma}_i)]^2} && (\varrho_p \text{ is a metric on } \mathbb{R})\\
&\le \sqrt{\sum_{i=1}^n[\varrho_p(\sigma_i,\widehat{\sigma}_i)]^2}+\sqrt{\sum_{i=1}^n[\varrho_p(\widehat{\sigma}_i,\widetilde{\sigma}_i)]^2}\\
&\le 2^{-1/p}\sqrt{\sum_{i=1}^n[\chi(\sigma_i,\widehat{\sigma}_i)]^2}+2^{-1/p}\sqrt{\sum_{i=1}^n[\chi(\widehat{\sigma}_i,\widetilde{\sigma}_i)]^2}\\
& && (\text{by Proposition 2.2})\\
&\le 2^{-1-1/p}\left(\|D_2^*-D_2^{-1}\|_{\rm F}+\|D_1^*-D_1^{-1}\|_{\rm F}\right) && (\text{by (6.2) and (6.3)}).
\end{aligned}$$

These inequalities complete the proof of Theorem 4.3. ∎

7. Generalized eigenvalue problems and generalized singular value problems. In this section, we discuss perturbations for *scaled generalized eigenvalue problems* and *scaled generalized singular value problems*. As we shall see, the results in previous sections, as well as those in Li [25], can be applied to derive relative perturbation bounds for these problems.

- *The generalized eigenvalue problem:*
  $A_1 - \lambda A_2 \equiv S_1^* H_1 S_1 - \lambda S_2^* H_2 S_2$ and $\widetilde{A}_1 - \lambda \widetilde{A}_2 \equiv S_1^* \widetilde{H}_1 S_1 - \lambda S_2^* \widetilde{H}_2 S_2$, where H_1 and H_2 are positive definite; $\|H_j^{-1}\|_2 \|\widetilde{H}_j - H_j\|_2 < 1$ for $j = 1, 2$; S_1 and S_2 are some square matrices and one of them is nonsingular.[7]
- *The generalized singular value problem:*
  $\{B_1, B_2\} \equiv \{G_1 S_1, G_2 S_2\}$ and $\{\widetilde{B}_1, \widetilde{B}_2\} \equiv \{\widetilde{G}_1 S_1, \widetilde{G}_2 S_2\}$, where G_1 and G_2 are nonsingular; $\|G_j^{-1}\|_2 \|\widetilde{G}_j - G_j\|_2 < 1$ for $j = 1, 2$; S_1 and S_2 are some square matrices and one of them is nonsingular.

For the scaled generalized eigenvalue problem just mentioned, without loss of generality, we consider the case when S_2 is nonsingular. Then the generalized eigenvalue problem for $A_1 - \lambda A_2 \equiv S_1^* H_1 S_1 - \lambda S_2^* H_2 S_2$ is equivalent to the standard eigenvalue problem for

$$A \stackrel{\text{def}}{=} H_2^{-1/2} S_2^{-*} S_1^* H_1 S_1 S_2^{-1} H_2^{-1/2},$$

and the generalized eigenvalue problem for $\widetilde{A}_1 - \lambda \widetilde{A}_2 \equiv S_1^* \widetilde{H}_1 S_1 - \lambda S_2^* \widetilde{H}_2 S_2$ is equivalent to the standard eigenvalue problem for

$$\widetilde{A} \stackrel{\text{def}}{=} D_2^* H_2^{-1/2} S_2^{-*} S_1^* \widetilde{H}_1 S_1 S_2^{-1} H_2^{-1/2} D_2,$$

where

$$D_2 = D_2^* \stackrel{\text{def}}{=} \left(I + H_2^{-1/2} (\Delta H_2) H_2^{-1/2}\right)^{-1/2} \quad \text{and} \quad \Delta H_2 \stackrel{\text{def}}{=} \widetilde{H}_2 - H_2.$$

So, bounding relative distances between the eigenvalues of $A_1 - \lambda A_2$ and those of $\widetilde{A}_1 - \lambda \widetilde{A}_2$ is transformed to bounding relative distances between the eigenvalues of A and those of $\widetilde{A}$. The latter can be accomplished in two steps:

1. Bounding relative distances between the eigenvalues of A and those of

$$\widehat{A} \stackrel{\text{def}}{=} D_2^* H_2^{-1/2} S_2^{-*} S_1^* H_1 S_1 S_2^{-1} H_2^{-1/2} D_2 = D_2^* A D_2.$$

2. Bounding relative distances between the eigenvalues of $\widehat{A}$ and those of $\widetilde{A}$.

Denote and order the eigenvalues of A, $\widehat{A}$, and $\widetilde{A}$ as

$$\lambda_1 \geq \cdots \geq \lambda_n, \quad \widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n, \quad \text{and} \quad \widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_n.$$

Set

$$D_1 = D_1^* \stackrel{\text{def}}{=} \left(I + H_1^{-1/2} (\Delta H_1) H_1^{-1/2}\right)^{-1/2} \quad \text{and} \quad \Delta H_1 \stackrel{\text{def}}{=} \widetilde{H}_1 - H_1.$$

By Theorem 3.1 on A and $\widehat{A} = D_2^* A D_2$, Theorem 3.2 on $\widehat{A} = X^* H_1 X$, and $\widetilde{A} = X^* \widetilde{H}_1 X$, where $X = S_1 S_2^{-1} H_2^{-1/2} D_2$, we have

$$\chi(\lambda_i, \widehat{\lambda}_i) \leq \|D_2 - D_2^{-1}\|_2 \quad \text{and} \quad \chi(\widehat{\lambda}_i, \widetilde{\lambda}_i) \leq \|D_1 - D_1^{-1}\|_2 \tag{7.1}$$

and

$$\sqrt{\sum_{i=1}^{n} \left[\chi(\lambda_i, \widehat{\lambda}_i)\right]^2} \leq \|D_2 - D_2^{-1}\|_{\mathrm{F}} \quad \text{and} \quad \sqrt{\sum_{i=1}^{n} \left[\chi(\widehat{\lambda}_i, \widetilde{\lambda}_i)\right]^2} \leq \|D_1 - D_1^{-1}\|_{\mathrm{F}}. \tag{7.2}$$

[7] When S_2 is singular, both pencils will have the same number of the eigenvalue $+\infty$. For convenience, we define the relative differences by any measure introduced in section 2 to be 0.

By Lemma 6.1, we have that if $\|D_1 - D_1^{-1}\|_2\|D_2 - D_2^{-1}\|_2 < 8$, then

$$\chi(\lambda_i, \widetilde{\lambda}_i) \le \frac{\chi(\lambda_i, \widehat{\lambda}_i) + \chi(\widehat{\lambda}_i, \widetilde{\lambda}_i)}{1 - \frac{1}{8}\chi(\lambda_i, \widehat{\lambda}_i)\chi(\widehat{\lambda}_i, \widetilde{\lambda}_i)} \le \frac{\|D_2 - D_2^{-1}\|_2 + \|D_1 - D_1^{-1}\|_2}{1 - \frac{1}{8}\|D_1 - D_1^{-1}\|_2\|D_2 - D_2^{-1}\|_2}$$

and

$$\begin{aligned}\sqrt{\sum_{i=1}^{n}\left[\chi(\lambda_i, \widetilde{\lambda}_i)\right]^2} &\le \sqrt{\sum_{i=1}^{n}\left[\frac{\chi(\lambda_i, \widehat{\lambda}_i) + \chi(\widehat{\lambda}_i, \widetilde{\lambda}_i)}{1 - \frac{1}{8}\chi(\lambda_i, \widehat{\lambda}_i)\chi(\widehat{\lambda}_i, \widetilde{\lambda}_i)}\right]^2} \\ &\le \frac{\sqrt{\sum_{i=1}^{n}\left[\chi(\lambda_i, \widehat{\lambda}_i)\right]^2} + \sqrt{\sum_{i=1}^{n}\left[\chi(\widehat{\lambda}_i, \widetilde{\lambda}_i)\right]^2}}{1 - \frac{1}{8}\max\limits_{1\le i\le n}\chi(\lambda_i, \widehat{\lambda}_i)\chi(\widehat{\lambda}_i, \widetilde{\lambda}_i)} \\ &\le \frac{\|D_2 - D_2^{-1}\|_{\mathrm{F}} + \|D_1 - D_1^{-1}\|_{\mathrm{F}}}{1 - \frac{1}{8}\|D_1 - D_1^{-1}\|_2\|D_2 - D_2^{-1}\|_2}.\end{aligned}$$

Notice also that for $j = 1, 2$ and for any unitarily invariant norm $|||\cdot|||$,

$$|||D_j - D_j^{-1}||| \le \frac{\|H_j^{-1}\|_2|||\Delta H_j|||}{\sqrt{1 - \|H_j^{-1}\|_2\|\Delta H_j\|_2}}.$$

So we have proved the following.

THEOREM 7.1. *Let $A_1 - \lambda A_2 \equiv S_1^* H_1 S_1 - \lambda S_2^* H_2 S_2$ and $\widetilde{A}_1 - \lambda\widetilde{A}_2 \equiv S_1^*\widetilde{H}_1 S_1 - \lambda S_2^*\widetilde{H}_2 S_2$, where H_1 and H_2 are $n \times n$, positive definite, and $\|H_j^{-1}\|_2\|\widetilde{H}_j - H_j\|_2 < 1$ for $j = 1, 2$. S_1 and S_2 are some square matrices and one of them is nonsingular. Let the generalized eigenvalues of $A_1 - \lambda A_2$ and $\widetilde{A}_1 - \lambda\widetilde{A}_2$ be*

$$\lambda_1 \ge \cdots \ge \lambda_n \quad \textit{and} \quad \widetilde{\lambda}_1 \ge \cdots \ge \widetilde{\lambda}_n.$$

If $\theta_1\theta_2\|\Delta H_1\|_2\|\Delta H_2\|_2 < 8$, then

$$\begin{aligned}\max_{1\le i\le n}\chi(\lambda_i, \widetilde{\lambda}_i) &\le \frac{\theta_1\|\Delta H_1\|_2 + \theta_2\|\Delta H_2\|_2}{1 - \frac{1}{8}\theta_1\theta_2\|\Delta H_1\|_2\|\Delta H_2\|_2}, \\ \sqrt{\sum_{i=1}^{n}\left[\chi(\lambda_i, \widetilde{\lambda}_i)\right]^2} &\le \frac{\theta_1\|\Delta H_1\|_{\mathrm{F}} + \theta_2\|\Delta H_2\|_{\mathrm{F}}}{1 - \frac{1}{8}\theta_1\theta_2\|\Delta H_1\|_2\|\Delta H_2\|_2},\end{aligned}$$

where $\theta_j \stackrel{\mathrm{def}}{=} \|H_j^{-1}\|_2 \Big/ \sqrt{1 - \|H_j^{-1}\|_2\|\Delta H_j\|_2}$ for $j = 1, 2$.

On the other hand, from (7.1), (7.2), and Proposition 2.2, we get

$$\varrho_p(\lambda_i, \widehat{\lambda}_i) \le 2^{-1/p}\|D_2 - D_2^{-1}\|_2 \quad \text{and} \quad \varrho_p(\widehat{\lambda}_i, \widetilde{\lambda}_i) \le 2^{-1/p}\|D_1 - D_1^{-1}\|_2$$

and

$$\sqrt{\sum_{i=1}^{n}\left[\varrho_p(\lambda_i, \widehat{\lambda}_i)\right]^2} \le 2^{-1/p}\|D_2 - D_2^{-1}\|_{\mathrm{F}} \quad \text{and} \quad \sqrt{\sum_{i=1}^{n}\left[\varrho_p(\widehat{\lambda}_i, \widetilde{\lambda}_i)\right]^2} \le 2^{-1/p}\|D_1 - D_1^{-1}\|_{\mathrm{F}}.$$

Since ϱ_p is a metric on $\mathbb{R}$, we have

$$\varrho_p(\lambda_i,\widetilde{\lambda}_i) \le \varrho_p(\lambda_i,\widehat{\lambda}_i) + \varrho_p(\widehat{\lambda}_i,\widetilde{\lambda}_i) \le 2^{-1/p}\left(\|D_2 - D_2^{-1}\|_2 + \|D_1 - D_1^{-1}\|_2\right)$$

and

$$\begin{aligned}\sqrt{\sum_{i=1}^n \left[\varrho_p(\lambda_i,\widetilde{\lambda}_i)\right]^2} &\le \sqrt{\sum_{i=1}^n \left[\varrho_p(\lambda_i,\widehat{\lambda}_i) + \varrho_p(\widehat{\lambda}_i,\widetilde{\lambda}_i)\right]^2}\\ &\le \sqrt{\sum_{i=1}^n \left[\varrho_p(\lambda_i,\widehat{\lambda}_i)\right]^2} + \sqrt{\sum_{i=1}^n \left[\varrho_p(\widehat{\lambda}_i,\widetilde{\lambda}_i)\right]^2}\\ &\le 2^{-1/p}\left(\|D_2 - D_2^{-1}\|_{\mathrm{F}} + \|D_1 - D_1^{-1}\|_{\mathrm{F}}\right).\end{aligned}$$

THEOREM 7.2. *Let all conditions of Theorem* 7.1*, except* $\|D_1 - D_1^{-1}\|_2\|D_2 - D_2^{-1}\|_2 < 8$*, which is no longer necessary, hold. Then*

$$\max_{1\le i\le n} \varrho_p(\lambda_i,\widetilde{\lambda}_i) \le 2^{-1/p}(\theta_1\|\Delta H_1\|_2 + \theta_2\|\Delta H_2\|_2),$$

$$\sqrt{\sum_{i=1}^n \left[\varrho_p(\lambda_i,\widetilde{\lambda}_i)\right]^2} \le 2^{-1/p}(\theta_1\|\Delta H_1\|_{\mathrm{F}} + \theta_2\|\Delta H_2\|_{\mathrm{F}}).$$

As to the scaled generalized singular value problem mentioned above, we shall consider instead its corresponding generalized eigenvalue problem [21, 36, 37] for

$$S_1^*G_1^*G_1S_1 - \lambda S_2^*G_2^*G_2S_2 \quad\text{and}\quad S_1^*\widetilde{G}_1^*\widetilde{G}_1S_1 - \lambda S_2^*\widetilde{G}_2^*\widetilde{G}_2S_2. \tag{7.3}$$

THEOREM 7.3. *Let* $\{B_1,B_2\} \equiv \{G_1S_1, G_2S_2\}$ *and* $\{\widetilde{B}_1,\widetilde{B}_2\} \equiv \{\widetilde{G}_1S_1,\widetilde{G}_2S_2\}$*, where* G_1 *and* G_2 *are* $n\times n$ *and nonsingular;* $\|G_j^{-1}\|_2\|\widetilde{G}_j - G_j\|_2 < 1$ *for* $j = 1,2$*;* S_1 *and* S_2 *are some square matrices and one of them is nonsingular. Let the generalized singular values of* $\{B_1,B_2\}$ *and* $\{\widetilde{B}_1,\widetilde{B}_2\}$ *be*

$$\sigma_1 \ge \cdots \ge \sigma_n \quad \textit{and} \quad \widetilde{\sigma}_1 \ge \cdots \ge \widetilde{\sigma}_n.$$

If $\delta_{12}\delta_{22} < 32$*, where*

$$\delta_{jt} = \left\|\left(I + (\Delta G_j)G_j^{-1}\right)^* - \left(I + (\Delta G_j)G_j^{-1}\right)^{-1}\right\|_t \quad \textit{for } j = 1,2 \textit{ and } t = 2,\mathrm{F},$$

then

$$\max_{1\le i\le n} \chi(\sigma_i,\widetilde{\sigma}_i) \le \frac{1}{2}\cdot\frac{\delta_{12}+\delta_{22}}{1-\frac{1}{32}\delta_{12}\delta_{22}},$$

$$\sqrt{\sum_{i=1}^n [\chi(\sigma_i,\widetilde{\sigma}_i)]^2} \le \frac{1}{2}\cdot\frac{\delta_{1\mathrm{F}}+\delta_{2\mathrm{F}}}{1-\frac{1}{32}\delta_{12}\delta_{22}}.$$

It can be proved that for $j = 1,2$ *and* $t = 2,\mathrm{F}$*,*

$$\begin{aligned}\delta_{jt} &\le \left(\frac{\|(\Delta G_j)G_j^{-1} + G_j^{-*}(\Delta G_j)^*\|_t}{\|(\Delta G_j)G_j^{-1}\|_t} + \frac{\|(\Delta G_j)G_j^{-1}\|_2}{1-\|(\Delta G_j)G_j^{-1}\|_2}\right)\|(\Delta G_j)G_j^{-1}\|_t\\ &\le \left(1 + \frac{1}{1-\|G_j^{-1}\|_2\|\Delta G_j\|_2}\right)\|G_j^{-1}\|_2\|\Delta G_j\|_t.\end{aligned}$$

Proof. Consider the case when S_2 is nonsingular. (The case when S_1 is nonsingular can be handled analogously.) By (7.3), we know that the singular values of $B \stackrel{\rm def}{=} G_1 S_1 S_2^{-1} G_2^{-1}$ and $\widetilde{B} \stackrel{\rm def}{=} \widetilde{G}_1 S_1 S_2^{-1} \widetilde{G}_2^{-1}$ are $\sigma_1 \geq \cdots \geq \sigma_n$ and $\widetilde{\sigma}_1 \geq \cdots \geq \widetilde{\sigma}_n$, respectively. Set

$$D_1 = I + (\Delta G_1) G_1^{-1},\ \Delta G_1 = \widetilde{G}_1 - G_1, \text{ and } D_2 = I + (\Delta G_2) G_2^{-1},\ \Delta G_2 = \widetilde{G}_2 - G_2;$$

then $\widetilde{B} = D_1 B D_2^{-1}$. By Theorem 4.1, we have

$$\max_{1\leq i\leq n} \chi(\sigma_i, \widetilde{\sigma}_i) \leq \frac{1}{2} \frac{\|D_1^* - D_1^{-1}\|_2 + \|D_2^{-*} - D_2\|_2}{1 - \frac{1}{32}\|D_1^* - D_1^{-1}\|_2 \|D_2^{-*} - D_2\|_2},$$

$$\sqrt{\sum_{i=1}^{n} [\chi(\sigma_i, \widetilde{\sigma}_i)]^2} \leq \frac{1}{2} \frac{\|D_1^* - D_1^{-1}\|_{\rm F} + \|D_2^{-*} - D_2\|_{\rm F}}{1 - \frac{1}{32}\|D_1^* - D_1^{-1}\|_2 \|D_2^{-*} - D_2\|_2},$$

as were to be shown. ∎

By the first half of the proof of Theorem 7.3 and by Theorem 4.3, we can prove the following.

THEOREM 7.4. *Let all conditions of Theorem* 7.3*, except* $\delta_{12}\delta_{22} < 32$*, which is no longer necessary, hold. Then*

$$\max_{1\leq i\leq n} \varrho_p(\sigma_i, \widetilde{\sigma}_i) \leq \frac{1}{2^{1+1/p}} (\delta_{12} + \delta_{22}),$$

$$\sqrt{\sum_{i=1}^{n} [\varrho_p(\sigma_i, \widetilde{\sigma}_i)]^2} \leq \frac{1}{2^{1+1/p}} (\delta_{1{\rm F}} + \delta_{2{\rm F}}).$$

8. Conclusions. We have developed a relative perturbation theory for eigenvalue and singular value variations under multiplicative perturbations. In the theory, extensions of the celebrated Hoffman–Wielandt and Weyl–Lidskii theorems from the classical perturbation theory are made. Our extensions use two kinds of relative distance: ϱ_p and χ. Topologically, these new relative distances are equivalent to the classical measurement (2.1) for relative accuracy, but the new distances have better mathematical properties. It is proved that ϱ_p is indeed a metric on $\mathbb{R}$ while χ is not. Often it is the case that perturbation bounds using χ are sharper than bounds using ϱ_p.

Our unifying treatment in this paper covers many previously studied cases and yields bounds that are at least as sharp as existing ones. Our results are applicable to the computations of sharp error bounds in the Demmel–Kahan QR [8] algorithm and the Fernando–Parlett implementation of the Rutishauser QD algorithm [14]; see Li [23].

Previous approaches to building a relative perturbation theory are more or less along the lines of using the min-max principle for Hermitian matrix eigenvalue problems. Our approach in this paper, however, is through deriving the perturbation equations (5.2) and (5.3). A major advantage of this new approach is that these perturbation equations will lead to the successful extensions in [25] of Davis–Kahan $\sin\theta$ theorems [5] and Wedin $\sin\theta$ theorems [38].

Appendix A. Proofs of Propositions 2.3 and 2.4.

LEMMA A.1. *Let $\alpha, \beta, \widetilde{\alpha}, \widetilde{\beta} \in \mathbb{R}$. If $\alpha \le \beta \le \widetilde{\beta} \le \widetilde{\alpha}$, then $\varrho_1(\alpha, \widetilde{\alpha}) \ge \varrho_1(\beta, \widetilde{\beta})$. If $\alpha \le \beta \le \widetilde{\beta} \le \widetilde{\alpha}$ and $\beta\widetilde{\beta} \ge 0$, then $\varrho_p(\alpha, \widetilde{\alpha}) \ge \varrho_p(\beta, \widetilde{\beta})$ for $p > 1$, and it is strict if either $\alpha < \beta$ or $\widetilde{\beta} < \widetilde{\alpha}$ holds.*

Proof. We consider function $f(\xi)$ defined by

$$f(\xi) \stackrel{\text{def}}{=} \frac{1-\xi}{\sqrt[p]{1+|\xi|^p}}, \quad \text{where } -1 \le \xi \le 1.$$

When $p = 1$,

$$f(\xi) = \begin{cases} 1, & \text{for } -1 \le \xi \le 0, \\ \frac{2}{1+\xi} - 1, & \text{for } 0 \le \xi \le 1, \end{cases}$$

so $f(\xi)$ decreases monotonically and decreases strictly monotonically for $0 \le \xi \le 1$. We are about to prove that when $p > 1$, function $f(\xi)$ so defined is strictly monotonically decreasing. This is true if $p = \infty$. When $1 < p < \infty$, set $h(\xi) \stackrel{\text{def}}{=} [f(\xi)]^p$ and $g(\xi) \stackrel{\text{def}}{=} [f(-\xi)]^p$. Since, for $0 < \xi < 1$,

$$h'(\xi) = -\frac{p(1-\xi)^{p-1}(1+\xi^{p-1})}{(1+\xi^p)^2} < 0 \quad \text{and} \quad g'(\xi) = \frac{p(1+\xi)^{p-1}(1-\xi^{p-1})}{(1+\xi^p)^2} > 0,$$

for $0 < \xi < 1$, $h(\xi)$ is strictly monotonically decreasing and $g(\xi)$ is strictly monotonically increasing. Thus function $f(\xi)$ is strictly monotonically decreasing for $p > 1$.

There are four cases to deal with. Assume that at least one of $\alpha \le \beta$ and $\widetilde{\beta} \le \widetilde{\alpha}$ is strict.

1. $0 \le \alpha \le \beta \le \widetilde{\beta} \le \widetilde{\alpha}$, then $0 \le \alpha/\widetilde{\alpha} < \beta/\widetilde{\beta} \le 1$; thus
$$\varrho_p(\alpha, \widetilde{\alpha}) = f(\alpha/\widetilde{\alpha}) > f(\beta/\widetilde{\beta}) = \varrho_p(\beta, \widetilde{\beta}).$$
2. $\alpha \le 0 \le \beta \le \widetilde{\beta} \le \widetilde{\alpha}$ or $\alpha \le \beta \le \widetilde{\beta} \le 0 \le \widetilde{\alpha}$; then
$$\varrho_p(\alpha, \widetilde{\alpha}) \ge 1 \ge \varrho_p(\beta, \widetilde{\beta}).$$
It is easy to verify that the equalities in the two inequality signs cannot be satisfied simultaneously.
3. $\alpha \le \beta \le 0 \le \widetilde{\beta} \le \widetilde{\alpha}$. Only $p = 1$ shall be considered:
$$\varrho_1(\alpha, \widetilde{\alpha}) = 1 = \varrho_1(\beta, \widetilde{\beta}).$$
4. $\alpha \le \beta \le \widetilde{\beta} \le \widetilde{\alpha} \le 0$, then $0 \le \widetilde{\alpha}/\alpha < \widetilde{\beta}/\beta \le 1$; thus
$$\varrho_p(\alpha, \widetilde{\alpha}) = f(\widetilde{\alpha}/\alpha) > f(\widetilde{\beta}/\beta) = \varrho_p(\beta, \widetilde{\beta}).$$

The proof is completed. ∎

Remark A.1. In Lemma A.1, assumption $\beta\widetilde{\beta} \ge 0$ for the case $p > 1$ is essential. A *counterexample* is the following: let $\xi > \zeta > 0$, and let $\alpha = -\zeta \le \beta = -\zeta < \widetilde{\beta} = \zeta < \widetilde{\alpha} < \xi$. Then

$$\varrho_p(\alpha, \widetilde{\alpha}) = \frac{\xi + \zeta}{\sqrt[p]{\xi^p + \zeta^p}} < 2^{1-1/p} = \varrho_p(\beta, \widetilde{\beta}).$$

Proof of Proposition 2.3. For any permutation τ of $\{1,2,\ldots,n\}$, the idea of our proof is to construct $n+1$ permutations τ_j such that

$$\tau_0 = \tau, \qquad \tau_n = \text{identity permutation},$$

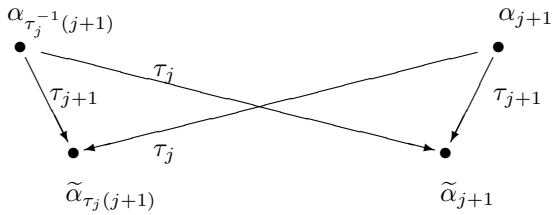
and for $j = 0, 1, 2, \ldots, n-1$,

$$\max_{1\le i\le n} \varrho_p(\alpha_i, \widetilde{\alpha}_{\tau_j(i)}) \ge \max_{1\le i\le n} \varrho_p(\alpha_i, \widetilde{\alpha}_{\tau_{j+1}(i)}).$$

The construction of these τ_j's goes as follows. Set $\tau_0 = \tau$. Given τ_j, if $\tau_j(j+1) = j+1$, set $\tau_{j+1} = \tau_j$; otherwise, define

$$\tau_{j+1}(i) = \begin{cases} \tau_j(i), & \text{for } \tau_j^{-1}(j+1) \ne i \ne j+1, \\ j+1, & \text{for } i = j+1, \\ \tau_j(j+1), & \text{for } i = \tau_j^{-1}(j+1). \end{cases}$$

In this latter case, τ_j and τ_{j+1} differ only at two indices as shown in the following picture (notice that $\tau_j^{-1}(j+1) > j+1$ and $\tau_j(j+1) > j+1$):



With Lemma A.1, it is easy to prove that

$$\begin{aligned} &\max\left\{\varrho_p(\alpha_{j+1}, \widetilde{\alpha}_{\tau_j(j+1)}),\ \varrho_p(\alpha_{\tau_j^{-1}(j+1)}, \widetilde{\alpha}_{j+1})\right\} \\ &\ge \max\left\{\varrho_p(\alpha_{j+1}, \widetilde{\alpha}_{j+1}),\ \varrho_p(\alpha_{\tau_j^{-1}(j+1)}, \widetilde{\alpha}_{\tau_j(j+1)})\right\}. \end{aligned}$$

Thus τ_j's so constructed have the desired properties. ∎

A proof of Proposition 2.4 can be given analogously with the help of the first inequality of Lemma 6.1 and the following lemma.

LEMMA A.2. *Let $\alpha_1 \ge \alpha_2 > 0$ and $\widetilde{\alpha}_1 \ge \widetilde{\alpha}_2 > 0$. Then*

$$[\chi(\alpha_1, \widetilde{\alpha}_1)]^2 + [\chi(\alpha_2, \widetilde{\alpha}_2)]^2 \le [\chi(\alpha_1, \widetilde{\alpha}_2)]^2 + [\chi(\alpha_2, \widetilde{\alpha}_1)]^2,$$

and the equality holds if and only if either $\alpha_1 = \alpha_2$ or $\widetilde{\alpha}_1 = \widetilde{\alpha}_2$.

Proof. It can be verified that

$$\begin{aligned} &\frac{(\widetilde{\alpha}_1 - \alpha_1)^2}{\widetilde{\alpha}_1\alpha_1} + \frac{(\widetilde{\alpha}_2 - \alpha_2)^2}{\widetilde{\alpha}_2\alpha_2} - \frac{(\widetilde{\alpha}_2 - \alpha_1)^2}{\widetilde{\alpha}_2\alpha_1} - \frac{(\widetilde{\alpha}_1 - \alpha_2)^2}{\widetilde{\alpha}_1\alpha_2} \\ &\qquad = -\frac{(\alpha_1 - \alpha_2)(\widetilde{\alpha}_1 - \widetilde{\alpha}_2)(\widetilde{\alpha}_1\widetilde{\alpha}_2 + \alpha_1\alpha_2)}{\widetilde{\alpha}_1\alpha_1\widetilde{\alpha}_2\alpha_2} \le 0, \end{aligned}$$

and the equality holds if and only if either $\alpha_1 = \alpha_2$ or $\widetilde{\alpha}_1 = \widetilde{\alpha}_2$. ∎

Appendix B. ϱ_p is a metric on $\mathbb{R}$. Throughout this appendix, we will be working with real numbers. The definition (2.2) of ϱ_p immediately implies that

1. $\varrho_p(\alpha, \widetilde{\alpha}) \ge 0$; and $\varrho_p(\alpha, \widetilde{\alpha}) = 0$ if and only if $\alpha = \widetilde{\alpha}$.

2. $\varrho_p(\alpha, \widetilde{\alpha}) = \varrho_p(\widetilde{\alpha}, \alpha)$.

So it remains to show that ϱ_p satisfies the triangle inequality

$$\varrho_p(\alpha, \gamma) \le \varrho_p(\alpha, \beta) + \varrho_p(\beta, \gamma) \quad \text{for } \alpha,\, \beta,\, \gamma \in \mathbb{R} \tag{B.1}$$

to conclude that the following holds.

THEOREM B.1. *ϱ_p is a metric on $\mathbb{R}$.*

We strongly conjecture that ϱ_p is a metric on $\mathbb{C}$. Unfortunately, we are unable to prove it at this point.

Since ϱ_p is symmetric with respect to its two arguments, we may assume, without loss of generality, that from now on

$$\alpha \le \gamma. \tag{B.2}$$

There are three possible positions for β:

$$\beta \le \alpha \quad \text{or} \quad \alpha < \beta \le \gamma \quad \text{or} \quad \gamma < \beta. \tag{B.3}$$

The hardest part of our proof is to show that (B.1) holds for the second position of β in (B.3). We state it in the following lemma whose proof is postponed to the end of this section.

LEMMA B.2. *Inequality* (B.1) *holds for $\alpha \le \beta \le \gamma$, and the equality holds if and only if $\beta = \alpha$ or $\beta = \gamma$.*

With this lemma, we are now ready to prove (B.1).

Proof of (B.1). The proof is divided into two different cases.

- *The case $\alpha\gamma \ge 0$.* Lemma B.2 says that (B.1) is true if $\alpha \le \beta \le \gamma$. If either $\beta < \alpha$ or $\gamma < \beta$, by Lemma A.1, we have

$$\varrho_p(\alpha, \gamma) \le \begin{cases} \varrho_p(\alpha, \beta) \le \varrho_p(\alpha, \beta) + \varrho_p(\beta, \gamma), & \text{if } \gamma \le \beta, \\ \varrho_p(\beta, \gamma) \le \varrho_p(\alpha, \beta) + \varrho_p(\beta, \gamma), & \text{if } \beta \le \alpha. \end{cases}$$

- *The case $\alpha\gamma < 0$.* We may assume $\alpha < 0$ and $\gamma > 0$ (see (B.2)). Consider the three possible positions (B.3) for β.
  1. $\beta \le \alpha < 0$. In this subcase, $1/\alpha \le 1/\beta < 0 < 1/\gamma$. By Lemma B.2, we have

$$\varrho_p(\alpha, \gamma) = \varrho_p(1/\alpha, 1/\gamma) \le \varrho_p(1/\alpha, 1/\beta) + \varrho_p(1/\beta, 1/\gamma) = \varrho_p(\alpha, \beta) + \varrho_p(\beta, \gamma).$$

  2. $\alpha \le \beta \le \gamma$. This subcase has been taken care of by Lemma B.2.
  3. $0 < \gamma \le \beta$. In this subcase, $1/\alpha < 0 < 1/\beta \le 1/\gamma$. The rest is the same as in subcase 1 above.

The proof is completed. □

Proof of Lemma B.2. Since both swapping α and γ and multiplying α, β, γ all by -1 lose no generality, we may further assume that

$$\alpha \le |\alpha| \le \gamma. \tag{B.4}$$

Inequality (B.1) clearly holds if one of α, β, γ is zero or if $\beta = \alpha$, $\beta = \gamma$, or $\alpha = \gamma$. So from now on we assume

$$\alpha < \beta < \gamma \quad \text{and} \quad \alpha \ne 0,\, \beta \ne 0,\, \gamma \ne 0.$$

For $1 \le p < \infty$,

$$\begin{aligned}
\varrho_p(\alpha,\gamma) &= \frac{\gamma-\alpha}{\sqrt[p]{\gamma^p+|\alpha|^p}} = \frac{\gamma-\beta+\beta-\alpha}{\sqrt[p]{\gamma^p+|\alpha|^p}} = \frac{\gamma-\beta}{\sqrt[p]{\gamma^p+|\alpha|^p}} + \frac{\beta-\alpha}{\sqrt[p]{\gamma^p+|\alpha|^p}} \\
&= \frac{\gamma-\beta}{\sqrt[p]{\gamma^p+|\beta|^p}} + \frac{\beta-\alpha}{\sqrt[p]{|\beta|^p+|\alpha|^p}} \\
&\quad +(\gamma-\beta)\left(\frac{1}{\sqrt[p]{\gamma^p+|\alpha|^p}} - \frac{1}{\sqrt[p]{\gamma^p+|\beta|^p}}\right) \\
&\quad +(\beta-\alpha)\left(\frac{1}{\sqrt[p]{\gamma^p+|\alpha|^p}} - \frac{1}{\sqrt[p]{|\alpha|^p+|\beta|^p}}\right) \\
&= \varrho_p(\alpha,\beta) + \varrho_p(\beta,\gamma) + h,
\end{aligned}$$

where

$$\begin{aligned}
h = &\frac{(\gamma-\beta)(|\beta|^p-|\alpha|^p)}{\sqrt[p]{\gamma^p+|\alpha|^p}\,\sqrt[p]{\gamma^p+|\beta|^p}} \cdot \frac{\sqrt[p]{\gamma^p+|\beta|^p}-\sqrt[p]{\gamma^p+|\alpha|^p}}{|\beta|^p-|\alpha|^p} \\
&+\frac{(\beta-\alpha)(|\beta|^p-\gamma^p)}{\sqrt[p]{\gamma^p+|\alpha|^p}\,\sqrt[p]{|\alpha|^p+|\beta|^p}} \cdot \frac{\sqrt[p]{|\alpha|^p+|\beta|^p}-\sqrt[p]{\gamma^p+|\alpha|^p}}{|\beta|^p-\gamma^p}.
\end{aligned}$$

The second factors of the two summands in h are always nonnegative. Now if $\alpha < \beta \le |\alpha| \le \gamma$, then $|\beta|^p - |\alpha|^p \le 0$ and $|\beta|^p - \gamma^p < 0$, and thus $h < 0$. Hence $\varrho_p(\alpha,\gamma) < \varrho_p(\alpha,\beta) + \varrho_p(\beta,\gamma)$. Consider now $|\alpha| < \beta < \gamma$. Then

$$\begin{aligned}
h = &\frac{(\gamma-\beta)(\beta-|\alpha|)}{\sqrt[p]{\gamma^p+|\alpha|^p}} \left(\frac{1}{\sqrt[p]{\gamma^p+\beta^p}} \cdot \frac{\beta^p-|\alpha|^p}{\beta-|\alpha|} \cdot \frac{\sqrt[p]{\gamma^p+\beta^p}-\sqrt[p]{\gamma^p+|\alpha|^p}}{\beta^p-|\alpha|^p}\right. \\
&\left. - \frac{1}{\sqrt[p]{|\alpha|^p+\beta^p}} \cdot \frac{\gamma^p-\beta^p}{\gamma-\beta} \cdot \frac{\sqrt[p]{|\alpha|^p+\beta^p}-\sqrt[p]{\gamma^p+|\alpha|^p}}{\beta^p-\gamma^p}\right) \\
< &\, 0.
\end{aligned}$$

The last inequality is true because $\sqrt[p]{\gamma^p+\beta^p} > \sqrt[p]{|\alpha|^p+\beta^p} \Rightarrow \frac{1}{\sqrt[p]{\gamma^p+\beta^p}} < \frac{1}{\sqrt[p]{|\alpha|^p+\beta^p}}$ and

$$\begin{aligned}
&0 < \frac{\beta^p-|\alpha|^p}{\beta-|\alpha|} \le \frac{\gamma^p-\beta^p}{\gamma-\beta}, \\
&0 < \frac{\sqrt[p]{\gamma^p+\beta^p}-\sqrt[p]{\gamma^p+|\alpha|^p}}{(\gamma^p+\beta^p)-(\gamma^p+|\alpha|^p)} \le \frac{\sqrt[p]{|\alpha|^p+\beta^p}-\sqrt[p]{\gamma^p+|\alpha|^p}}{(|\alpha|^p+\beta^p)-(\gamma^p+|\alpha|^p)}
\end{aligned}$$

by Lemma B.3, since for $1 < p < \infty$, $f(x) = x^p$ is convex and $g(x) = \sqrt[p]{x}$ is concave. So we also have $\varrho_p(\alpha,\gamma) < \varrho_p(\alpha,\beta) + \varrho_p(\beta,\gamma)$ for $|\alpha| < \beta < \gamma$. The proof for the case $p < \infty$ is completed.

When $p = \infty$, (B.4) and $\alpha < \beta < \gamma$ imply $|\gamma| > \max\{|\alpha|, |\beta|\}$. So

$$\begin{aligned}
\varrho_\infty(\alpha,\gamma) &= \frac{\gamma-\alpha}{\gamma} = \frac{\gamma-\beta}{\gamma} + \frac{\beta-\alpha}{\gamma} \\
&= \frac{\gamma-\beta}{\gamma} + \frac{\beta-\alpha}{\max\{|\alpha|,|\beta|\}} + (\beta-\alpha)\left(\frac{1}{\gamma} - \frac{1}{\max\{|\alpha|,|\beta|\}}\right) \\
&< \varrho_\infty(\alpha,\beta) + \varrho_\infty(\beta,\gamma),
\end{aligned}$$

as was to be shown. □

LEMMA B.3. *Suppose functions $f(x)$ and $g(x)$ are defined on the interval $[a, b]$, and suppose $f(x)$ is convex and $g(x)$ concave. Let $x, y, z \in [a, b]$ and $x \le y \le z$. Then*

$$\frac{f(y) - f(x)}{y - x} \le \frac{f(z) - f(y)}{z - y} \quad \textit{and} \quad \frac{g(y) - g(x)}{y - x} \ge \frac{g(z) - g(y)}{z - y}.$$

A proof of this lemma can be found in most mathematical analysis books; see, e.g., [31, section 1.4.4].

Acknowledgments. I thank Professor W. Kahan for his consistent encouragement and support, Professor J. Demmel for helpful discussions on open problems in this research area, and Professor B. N. Parlett for drawing my attention to Ostrowski's theorem. Thanks also go to Professor I. C. F. Ipsen for sending me the report [13]. Professor R. Bhatia and the referees' constructive comments, which improve the presentation considerably, are greatly appreciated.

REFERENCES

[1] J. BARLOW AND J. DEMMEL, *Computing accurate eigensystems of scaled diagonally dominant matrices*, SIAM J. Numer. Anal., 27 (1990), pp. 762–791.
[2] F. L. BAUER AND C. T. FIKE, *Norms and exclusion theorems*, Numer. Math., 2 (1960), pp. 137–141.
[3] R. BHATIA, *Matrix Analysis*, Graduate Texts in Mathematics 169, Springer-Verlag, New York, 1996.
[4] G. D. BIRKHOFF, *Tres observaciones sobre el algebra lineal*, Univ. Nac. de Tucuman Rev., Ser. A, 5 (1946), pp. 147–151.
[5] C. DAVIS AND W. KAHAN, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.
[6] P. DEIFT, J. DEMMEL, L.-C. LI, AND C. TOMEI, *The bidiagonal singular value decomposition and Hamiltonian mechanics*, SIAM J. Numer. Anal., 28 (1991), pp. 1463–1516.
[7] J. DEMMEL AND W. GRAGG, *On computing accurate singular values and eigenvalues of matrices with acyclic graphs*, Linear Algebra Appl., 185 (1993), pp. 203–217.
[8] J. DEMMEL AND W. KAHAN, *Accurate singular values of bidiagonal matrices*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 873–912.
[9] J. DEMMEL AND K. VESELIĆ, *Jacobi's method is more accurate than QR*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 1204–1245.
[10] G. DILENA, R. I. PELUSO, AND G. PIAZZA, *Results on the relative perturbation of the singular values of a matrix*, BIT, 33 (1993), pp. 647–653.
[11] S. C. EISENSTAT AND I. C. F. IPSEN, *Relative perturbation bounds for eigenspaces and singular vector subspaces*, in Proceedings of the Fifth SIAM Conference on Applied Linear Algebra, J. G. Lewis, ed., SIAM, Philadelphia, PA, 1994, pp. 62–66.
[12] S. C. EISENSTAT AND I. C. F. IPSEN, *Relative perturbation techniques for singular value problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1972–1988.
[13] S. C. EISENSTAT AND I. C. F. IPSEN, *Relative Perturbation Results for Eigenvalues and Eigenvectors of Diagonalisable Matrices*, Technical Report CRSC-TR96-6, Department of Mathematics, North Carolina State University, Raleigh, NC, 1996.
[14] K. V. FERNANDO AND B. N. PARLETT, *Accurate singular values and differential qd algorithms*, Numer. Math., 67 (1994), pp. 191–229.
[15] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 2nd ed., Johns Hopkins University Press, Baltimore, MD, 1989.
[16] M. GU AND S. C. EISENSTAT, *Relative Perturbation Theory for Eigenproblems*, Research Report YALEU/DCS/RR–934, Department of Computer Science, Yale University, New Haven, CT, 1993.
[17] A. J. HOFFMAN AND H. W. WIELANDT, *The variation of the spectrum of a normal matrix*, Duke Math. J., 20 (1953), pp. 37–39.
[18] R. A. HORN AND C. R. JOHNSON, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[19] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.
[20] W. Kahan, *Accurate Eigenvalues of a Symmetric Tridiagonal Matrix*, Technical Report CS41, Computer Science Department, Stanford University, Stanford, CA, 1966 (revised June 1968).
[21] R.-C. Li, *Bounds on perturbations of generalized singular values and of associated subspaces*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 195–234.
[22] R.-C. Li, *Norms of certain matrices with applications to variations of the spectra of matrices and matrix pencils*, Linear Algebra Appl., 182 (1993), pp. 199–234.
[23] R.-C. Li, *On Deflating Bidiagonal Matrices*, manuscript, Department of Mathematics, University of California, Berkeley, CA, 1994.
[24] R.-C. Li, *Relative Perturbation Theory:* (I) *Eigenvalue and Singular Value Variations*, Technical Report UCB//CSD-94-855, Computer Science Division, Department of EECS, University of California at Berkeley, 1994; LAPACK working note 85 (revised January 1996) available online at http://www.netlib.org/lapack/lawns/lawn84.ps
[25] R.-C. Li, *Relative Perturbation Theory:* (II) *Eigenspace and Singular Subspace Variations*, Technical Report UCB//CSD-94-856, Computer Science Division, Department of EECS, University of California at Berkeley, 1994; LAPACK working note 85 (revised January 1996 and April 1996), available at http://www.netlib.org/lapack/lawns/lawn85.ps.
[26] R.-C. Li, *Relative Perturbation Theory:* (III) *More Bounds on Eigenvalue Variation*, Linear Algebra Appl., 266 (1996), pp. 337–345.
[27] V. B. Lidskii, *The proper values of the sum and product of symmetric matrices*, Dokl. Akad. Nauk SSSR, 75 (1950), pp. 769–772 (in Russian). Translation by C. Benster available from the National Translation Center of the Library of Congress.
[28] R. Mathias, *Spectral perturbation bounds for positive definite matrices*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 959–980.
[29] R. Mathias and G. W. Stewart, *A block QR algorithm and the singular value decomposition*, Linear Algebra Appl., 182 (1993), pp. 91–100.
[30] L. Mirsky, *Symmetric gauge functions and unitarily invariant norms*, Quart. J. Math., 11 (1960), pp. 50–59.
[31] D. S. Mitrinovic, *Analytic Inequalities*, Springer-Verlag, New York, 1970.
[32] A. M. Ostrowski, *A quantitative formulation of Sylvester's law of inertia*, Proc. Nat. Acad. Sci. USA, 45 (1959), pp. 740–744.
[33] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[34] I. Slapničar, *Accurate Symmetric Eigenreduction by a Jacobi Method*, Ph.D. thesis, Fernuniversität–Gesamthochschule–Hagen, Fachbereich Mathematik, 1992.
[35] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, Boston, 1990.
[36] J.-G. Sun, *Perturbation analysis for the generalized singular value decomposition*, SIAM J. Numer. Anal., 20 (1983), pp. 611–625.
[37] C. F. Van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal., 13 (1976), pp. 76–83.
[38] P.-Å. Wedin, *Perturbation bounds in connection with singular value decomposition*, BIT, 12 (1972), pp. 99–111.
[39] H. Wielandt, *An extremum property of sums of eigenvalues*, Proc. Amer. Math. Soc., 6 (1955), pp. 106–110.
[40] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 983–1004, October 1998

010

ANALYSES, DEVELOPMENT, AND APPLICATIONS OF TLS ALGORITHMS IN FREQUENCY DOMAIN SYSTEM IDENTIFICATION*

RIK PINTELON†, PATRICK GUILLAUME‡, GERD VANDERSTEEN†, AND YVES ROLAIN†

Abstract. This paper gives an overview of frequency domain total least squares (TLS) estimators for rational transfer function models of linear time-invariant multivariable systems. The statistical performance of the different approaches are analyzed through their equivalent cost functions. Both generalized and bootstrapped total least squares (GTLS and BTLS) methods require the exact knowledge of the noise covariance matrix. The paper also studies the asymptotic (the number of data points going to infinity) behavior of the GTLS and BTLS estimators when the exact noise covariance matrix is replaced by the sample noise covariance matrix obtained from a (small) number of independent data sets. Even if only two independent repeated observations are available, it is shown that the estimates are still strongly consistent without any increase in the asymptotic uncertainty.

Key word. system identification

AMS subject classification. 93E12

PII. S0895479896309074

1. Introduction.

Total least squares (TLS) techniques have been applied with success to a wide variety of problems [26]. This paper gives an overview of its application to frequency domain identification of linear time-invariant multivariable systems. The key analysis tool used throughout the paper is the equivalent cost function minimized by the TLS method. Analyzing the cost function reveals the statistical properties of the TLS estimator, shows its shortcomings, and allows us, by comparison with the maximum likelihood (ML) approach, to propose weighted TLS versions with nearly ML properties.

The paper starts by defining the parametric model and the stochastic framework (section 2). Next, the TLS estimation of the model parameters is handled (section 3) and some extensions are given (section 4). Sections 5 and 6 study the properties of the generalized and bootstrapped total least squares (GTLS and BTLS) estimators when the true noise covariance matrix is replaced by the noise sample covariance matrix. The theory is illustrated in sections 7 and 8 by simulation and real measurement examples.

2. Multiple input, multiple output systems.

2.1. Model equations.

Consider a real, linear, time-invariant multivariable system without time delay with nu inputs and ny outputs. Assume that the input signals are periodic or time limited. The (discrete) Fourier spectra $U_0(j\omega)$ and

*Received by the editors September 9, 1996; accepted for publication (in revised form) by S. Van Huffel April 30, 1997; published electronically July 9, 1998. This research was supported by the Belgian National Fund for Scientific Research, the Flemish Community (Concerted Action IMMI), and the Belgian Government as part of the Belgian program on Inter-University Poles of Attraction (IUAP 4/2) initiated by the Belgian State, Prime Minister's Office, Science Policy Programming.
http://www.siam.org/journals/simax/19-4/30907.html

†Department ELEC, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium (Rik.Pintelon@vub.ac.be, gvanders@vub.ac.be, Yves.Rolain@vub.ac.be).

‡Department WERK, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium (paguilla@vub.ac.be).

$Y_0(j\omega)$ of, respectively, the input $u_0(t) = [u_{01}(t)u_{02}(t), \ldots, u_{0nu}(t)]^T$ and output $y_0(t) = [y_{01}(t)y_{02}(t), \ldots, y_{0ny}(t)]^T$ signals are related to each other through a transfer function matrix $G_0(j\omega) \in \mathcal{C}^{ny\times nu}$,

$$Y_0(j\omega) = G_0(j\omega)U_0(j\omega). \tag{1}$$

The (discrete) Fourier spectra $U_0(j\omega)$, $Y_0(j\omega)$ are primarily calculated from the knowledge of N samples of the measured time signals. Sometimes the (discrete) Fourier spectra are directly measured, for example, in high frequency network analyzers. If the input is periodic and an integer number of periods of the steady state response is measured, then the (discrete) Fourier spectra can be calculated without systematic errors through the discrete Fourier transforms (DFT) of the samples $u_0(nT_s)$ and $y_0(nT_s)$, $n = 0, 1, \ldots, N-1$, with T_s the sampling period [4]. If the input signal is time limited, then, by an appropriate choice of the measurement time, the cutoff frequency of the anti-alias filters, and the sampling frequency, the spectral leakage and alias errors of the DFT can be made arbitrarily small [4]. Equation (1) can then be evaluated at the excited DFT angular frequencies $\{\omega_1, \omega_2, \ldots, \omega_F\}$ with $\omega_k \in \{2\pi r/(NT_s)$; $r = 0, 1, \ldots, N/2\}$.

Unless $nu = 1$, it is impossible to calculate $G_0(j\omega)$ from (1) ($U_0(j\omega) \in \mathcal{C}^{nu}$ and $Y_0(j\omega) \in \mathcal{C}^{ny}$). Therefore the multiple input, multiple output (MIMO) experiment is often repeated M times with different excitation signals $U_0^{(i)}(j\omega) \in \mathcal{C}^{nu}$, $i = 1, 2, \ldots, M$. Denoting the corresponding output signals by $Y_0^{(i)}(j\omega) \in \mathcal{C}^{ny}$, $i = 1, 2, \ldots, M$, and redefining the input/output Fourier data as

$$\begin{aligned} Y_0(j\omega) &= \left[Y_0^{(1)}(j\omega)Y_0^{(2)}(j\omega), \ldots, Y_0^{(M)}(j\omega)\right] \in \mathcal{C}^{ny\times M}, \\ U_0(j\omega) &= \left[U_0^{(1)}(j\omega)U_0^{(2)}(j\omega), \ldots, U_0^{(M)}(j\omega)\right] \in \mathcal{C}^{nu\times M}, \end{aligned} \tag{2}$$

it can easily be seen that relationship (1) is still valid. If the rank of $U_0(j\omega)$ equals nu, then $U_0(j\omega)$ is regular and $G_0(j\omega) = Y_0(j\omega)U_0^+(j\omega)$, where superscript + denotes the Moore–Penrose pseudo-inverse [2]. Proceeding in this way, measurements of the true transfer function matrix $G_0(j\omega)$ can be obtained experimentally [10]. An easy quality check (= model validation) of the estimated parametric model consisting of comparing it to the measured transfer function matrix is then possible.

There exist many parametrizations of the transfer function matrix; for example, the state space representation, the matrix and the partial fraction descriptions, etc. [12]. The TLS approach requires a parametrization that leads to a model equation which is linear in the model parameters. When using the input and output Fourier matrices $U_0(j\omega)$ and $Y_0(j\omega)$ as primary data, then the left matrix fraction description is the only parametrization which results in linear relationship between the matrix coefficients. The left matrix fraction description writes the transfer function matrix as the ratio of two matrix polynomials

$$G(\Omega, X) = D^{-1}(\Omega, X)N(\Omega, X) = \left[\sum_{k=0}^{od} D_k\Omega^{od-k}\right]^{-1}\left[\sum_{k=0}^{on} N_k\Omega^{on-k}\right]. \tag{3}$$

$X = [D_0D_1, \ldots, D_{od} \quad N_0N_1, \ldots, N_{on}]^T \in \mathcal{R}^{n\times d}(n = (od+1)ny + (on+1)nu, \quad d = ny)$ are the model parameters, $N_k \in \mathcal{R}^{ny\times nu}$, $D_k \in \mathcal{R}^{ny\times ny}$ are the numerator and denominator real matrix coefficients, and Ω is a generalized frequency variable. The

generalized frequency variable Ω equals $j\omega$ for continuous time systems, $\exp(-j\omega T_s)$ for discrete time systems, $\tanh(j\omega\tau)$ for commensurate microwave systems, and $\sqrt{j\omega}$ for diffusion phenomena.

Notice that the transfer function model (3) is not identifiable since $G(\Omega, X\Lambda) = G(\Omega, X)$ for any regular matrix $\Lambda \in \mathcal{R}^{d\times d}$. To remove the parameter redundancy, parameter constraints have to be imposed. Model (3) can be made identifiable by fixing one matrix coefficient of the denominator polynomial, e.g., $D_{od} = I_d$, or by imposing a 2-norm constraint on the parameter matrix X, i.e., $X^T X = I_d$.

Using (1), (2), and (3), the model equation is readily obtained.

$$D(\Omega, X)Y_0(j\omega) - N(\Omega, X)U_0(j\omega) = 0. \tag{4}$$

Since (4) is linear in the model parameters X and in the input/output Fourier data $Z_0^T(j\omega_k) = [Y_0^T(j\omega_k), U_0^T(j\omega_k)]$ ($Z_0 \in \mathcal{C}^{(ny+nu)\times M}$), it can be reformulated as

$$\begin{aligned} &X^T S(j\omega_k) Z_0(j\omega_k) = 0 \\ &\text{where } S(j\omega) = \text{block diag}([\Omega^{od}, \ldots, 1]^T \otimes I_{ny}, -[\Omega^{on}, \ldots, 1]^T \otimes I_{nu}), \end{aligned} \tag{5}$$

with $\otimes$ the Kronecker product [3]. Rewriting equation (5), evaluated at the considered F frequencies as an overdetermined set of $2F$ real-valued linear equations, gives

$$\begin{aligned} &A_0 X = 0 \\ &\text{where } A_0 \in \mathcal{R}^{m\times n}, X \in \mathcal{R}^{n\times d}, m = 2F, n = (od+1)ny + (on+1)nu, d = ny, \end{aligned} \tag{6}$$

with $A_0 = [\text{real}(\alpha_{0\ 1}), \text{real}(\alpha_{0\ 2}), \ldots, \text{real}(\alpha_{0F}), \text{imag}(\alpha_{0\ 1}), \ldots, \text{imag}(\alpha_{0F})]^T$ and

$$\begin{aligned} \alpha_{0k}^T &= [Y_0^T(j\omega_k), U_0^T(j\omega_k)] S^T(j\omega_k) \\ &= [Y_0^T(j\omega_k)\Omega_k^{od}, Y_0^T(j\omega_k)\Omega_k^{od-1}, \ldots, Y_0^T(j\omega_k), -U_0^T(j\omega_k)\Omega_k^{on}, \\ &\quad -U_0^T(j\omega_k)\Omega_k^{on-1}, \ldots, -U_0^T(j\omega_k)]. \end{aligned} \tag{7}$$

The identification problem to be solved is finding an X of full column rank such that (6) is satisfied.

2.2. Stochastic framework. In practice, the model parameters X are estimated using noisy measurements $U(j\omega_k)$, $Y(j\omega_k)$ of the true (deterministic) input and output DFT spectra $U_0(j\omega_k)$, $Y_0(j\omega_k)$. By introducing $Z^T(j\omega_k) = [Y^T(j\omega_k), U^T(j\omega_k)]$ ($Z \in \mathcal{C}^{(ny+nu)\times M}$), the errors-in-variables equations are

$$Z(j\omega_k) = Z_0(j\omega_k) + \Delta Z(j\omega_k), \tag{8}$$

with $Z_0(j\omega_k)$ the true (unknown) values and $\Delta Z(j\omega_k)$ the errors. Relying on the properties of the discrete Fourier transform [5], it is reasonable to make the following assumption.

ASSUMPTION 2.1. *$\Delta Z(j\omega_k)$, $k = 1, 2, \ldots, F$ are zero-mean, mixing*[1] *(over the frequency), complex distributed random matrices with known Hermitian-symmetric*

[1] Intuitively this means that the (frequency) span of the dependency is limited or the correlation of the errors over the frequency must tend sufficiently fast to zero. See [5] for a formal definition. For example, filtered, white noise (with bounded moments) is mixing.

matrices

$$E\{\Delta Z(j\omega_k)\Delta Z^H(j\omega_k)\} = C_Z(j\omega_k) = \begin{bmatrix} C_Y(j\omega_k) & C_{YU}(j\omega_k) \\ C_{YU}^H(j\omega_k) & C_U(j\omega_k) \end{bmatrix},$$

satisfying $E\{\Delta Z(j\omega_k)\Delta Z^T(j\omega_k)\} = 0$.

It is also reasonable to assume that the experimental conditions of the repeated MIMO measurements are such that the columns of $\Delta Z(j\omega_k)$ are independent and identically distributed [10]. These restrictions are, however, not included in Assumption 2.1; the columns of $\Delta Z(j\omega_k)$ may be correlated and may have different covariance matrices. In section 3.2, it will be shown that the GTLS solutions require neither the knowledge of the individual covariance matrices of the columns of $\Delta Z(j\omega_k)$ nor the correlation of $\Delta Z(j\omega_k)$ over the frequency. The optimally weighted GTLS solutions, however, need this information. Therefore the following additional assumption is made.

ASSUMPTION 2.2. *The errors $\Delta Z(j\omega_k)$ are independent over the frequency; the columns of $\Delta Z(j\omega_k)$ are independent and complex normally distributed random vectors.*

Putting the noisy values (8) into model equation (6) defines the noisy matrix A

$$A = A_0 + \Delta A \tag{9}$$

with

$$A = [\text{real}(\alpha_1), \text{real}(\alpha_2), \ldots, \text{real}(\alpha_F), \text{imag}(\alpha_1), \ldots, \text{imag}(\alpha_F)]^T, \tag{10}$$

$$\alpha_k^T = Z^T(j\omega_k)S^T(j\omega_k). \tag{11}$$

Using the stochastic errors-in-variables framework (9), the identification problem is reformulated as follows. Find an X of full column rank (satisfying $X^TX = I_d$) such that AX is "as small as possible" (in Frobenius norm).

3. Parameter estimation. The weighted generalized total least squares (WGTLS) solution to the estimation problem $AX = 0$ is [26]

$$\arg\min_{\hat{A},X} \|W(A - \hat{A})C^{-1}\|_F^2 \quad \text{subject to } \hat{A}X = 0 \quad \text{and } X^TX = I_d. \tag{12}$$

$W \in \mathcal{R}^{m\times m}$ is a left weighting matrix and $C \in \mathcal{R}^{n\times n}$ is a square root of the column covariance matrix of WA: $C^TC = E\{\Delta A^TW^TW\Delta A\}$. The matrix C is singular for identification problems with singular covariance matrices $C_Z(j\omega_k)$ $\forall k$. This occurs when one or more DFT spectra are noise free or when some of the noise sources are totally correlated (see, for example, section 4.1). Elimination of $\hat{A}$ in (12) gives the equivalent cost function minimized by the WGTLS estimator (see Appendix 1)

$$\arg\min_X \text{ trace}((WAX)[X^TC^TCX]^{-1}(WAX)^T) \quad \text{subject to } X^TX = I_d. \tag{13}$$

Under Assumption 2.1, and provided some regularity conditions are satisfied (see section 5.2), it can be proven that the WGTLS solution X_{WGTLS} for a deterministic weighting W of full rank is strongly consistent and asymptotical normally distributed

($m \to \infty$) [23]. The solution X_{WGTLS} is *not* calculated by minimizing cost function (13), but through the generalized singular value decomposition (GSVD) of the matrix pair $(WA,\ C)$ [1], [13].

The multidimensional TLS problem (12) ($d = ny$) can be reformulated into a one-dimensional problem ($d = 1$). Indeed, applying the vec operator to $X^T A^T \approx 0$ gives $(A \otimes I_d)v \approx 0$ with $v = \mathrm{vec}(X^T)$. The one-dimensional constraint $v^T v = 1$ is, however, not sufficient to remove the parameter redundancy of a left matrix fraction description when $ny > 1$. That problem can be circumvented by using, for instance, another parametrization, e.g., a common denominator model $D_k = d_k I_d$ ($d_k \in \mathcal{R}, k = 0, 1, \ldots, od$), or by fixing a matrix coefficient of the denominator polynomial, e.g., $D_{od} = I_d$. This leads to a parameter vector of reduced dimension satisfying $\mathrm{vec}(X^T) = Lx$ with L a constant regular matrix of full column rank (see Appendix 2). The constraint $x^T x = 1$ is now sufficient to remove the parameter redundancy. The one-dimensional weighted total least squares (WTLS) solution to the estimation problem is

$$\begin{aligned} &\arg\min_{\hat{\mathcal{A}},x} \|W(\mathcal{A} - \hat{\mathcal{A}})C^{-1}\|_F^2 \quad \text{with } \mathcal{A} = (A \otimes I_{ny})L \\ &\text{subject to } \hat{\mathcal{A}}x = 0 \quad \text{and } x^T x = 1. \end{aligned} \tag{14}$$

$W \in \mathcal{R}^{(m \cdot ny) \times (m \cdot ny)}$ is a left weighting matrix and $C \in \mathcal{R}^{(n \cdot ny) \times (n \cdot ny)}$ is a square root of the column covariance matrix of $W\mathcal{A}$: $C^T C = E\{\Delta\mathcal{A}^T W^T W \Delta\mathcal{A}\}$. Elimination of $\hat{\mathcal{A}}$ in (14) results after some calculations using Kronecker algebra [3] in the following equivalent cost function (replace A and X by, respectively, $\mathcal{A}$ and x in (13))

$$\begin{aligned} &\arg\min_x \left(\frac{(W\mathrm{vec}((AX)^T))^T W\ \mathrm{vec}((AX)^T)}{x^T C^T C x} \right) \\ &\text{subject to } \mathrm{vec}(X^T) = Lx \quad \text{and } x^T x = 1. \end{aligned} \tag{15}$$

The one-dimensional WGTLS solution x_{WGTLS} is calculated through the GSVD of the matrix pair $(W\mathcal{A}, C)$. Its usefulness will become clear when discussing the optimal weighting of GTLS estimators (see section 3.4).

Note that the one-dimensional WGTLS solution (15) does not exploit the special structure of $\mathcal{A} = (A \otimes I_{ny})L$. When noise and/or modeling errors[2] are present, it will be different from the exact one-dimensional solution and hence also from the original multidimensional estimates. Indeed, the exact solution satisfies $(\hat{A} \otimes I_{ny})L\hat{x} = 0$, where $\hat{A} \otimes I_{ny}$ is of rank $ny(n - ny)$, while the WGTLS solution satisfies $\hat{\mathcal{A}}\hat{x} = 0$, where $\hat{\mathcal{A}}$ is of rank $n \cdot ny - 1$. The exact one-dimensional solution can be calculated using the structured TLS method [25]. It has a smaller uncertainty than (15) but requires an iterative algorithm. Generalizations, refinements, and convergence issues of this algorithm are currently under study [25].

It is readily verified that the *multidimensional* WGTLS cost function (13) is scale invariant: replacing X by $X\Lambda$ in (13), with $\Lambda \in \mathcal{R}^{d \times d}$ a regular scaling matrix, gives the same cost function. The *one-dimensional* WGTLS cost function (15) is *not* scale invariant with respect to (w.r.t.) Λ: replacing $\mathrm{vec}(X^T)$ by $\mathrm{vec}(\Lambda^T X^T)$ in (15) yields another cost function. Consequently, the one-dimensional WGTLS estimates will depend on the particular choice of L (i.e., the parameter constraint).

[2]Model errors occur when the polynomial orders *od* and/or *on* in (3) are too small and/or when nonlinear distortions are present.

The one-dimensional WGTLS cost function (15) is, however, scale invariant w.r.t. $\Lambda = \lambda I_{ny} (\lambda \neq 0)$, which means that the corresponding estimates are independent of the particular constraint on x (e.g., $x_i = 1$ or $x^T x = 1$) [14].

3.1. TLS. Putting $W = I_m$, $C = I_n$ in (13) gives the *multidimensional* TLS estimates

$$V_{TLS}(X,Z) = \sum_{k=1}^{F} \operatorname{trace}(\varepsilon^H(j\omega_k, X)[X^T X]^{-1}\varepsilon(j\omega_k, X)) \quad \text{subject to } X^T X = I_d, \tag{16}$$

with $\varepsilon(j\omega_k, X) \in \mathcal{C}^{ny \times M}$ a matrix polynomial in Ω_k given by

$$\varepsilon(j\omega_k, X) = X^T \alpha_k = D(\Omega_k, X) Y(j\omega_k) - N(\Omega_k, X) U(j\omega_k). \tag{17}$$

Calculating the expected value of (16) gives

$$E\{V_{TLS}(X,Z)\} = V_{TLS}(X, Z_0) + \sum_{k=1}^{F} \operatorname{trace}([X^T X]^{-1} W_{ML}^{-2}(j\omega_k, X)), \tag{18}$$

where

$$\begin{aligned} W_{ML}^{-2}(j\omega_k, X) &= E\{\Delta\varepsilon(j\omega_k, X)\Delta\varepsilon^H(j\omega_k, X)\} = X^T S(j\omega_k) C_Z(j\omega_k) S^H(j\omega_k) X \\ &= (D(\Omega_k, X) C_Y(j\omega_k) D^H(\Omega_k, X) + N(\Omega_k, X) C_U(j\omega_k) N^H(\Omega_k, X) \\ &\quad -2\operatorname{herm}(D(\Omega_k, X) C_{YU}(j\omega_k) N^H(\Omega_k, X))), \end{aligned} \tag{19}$$

and with $\operatorname{herm}(A) = (A + A^H)/2$. Under Assumption 2.1, the TLS cost function (16) converges with probability one to its expected value (18) (for the proof see [23]). The first term on the right-hand side of (18) is minimal in the true parameter values ($V_{TLS}(X_0, Z_0) = 0$). However, since the second term is X-dependent, the expected value of the cost function is in general not minimal in the true model parameters X_0. As a consequence, the TLS estimate X_{TLS} is inconsistent. The residuals of the model equation $\varepsilon(j\omega_k, X)$ have a frequency independent weighting $X^T X = I_d$ in (16), which explains why the TLS estimates overemphasize the high frequency errors [11], [14].

Putting $W = I_{m \cdot ny}$, $C = I_{n \cdot ny}$ in (15) gives the *one-dimensional* TLS estimates

$$V_{TLS}(x,Z) = \sum_{k=1}^{F} \frac{\operatorname{trace}(\varepsilon^H(j\omega_k, X)\varepsilon(j\omega_k, X))}{x^T x} \quad \text{subject to } \operatorname{vec}(X^T) = Lx \quad \text{and} \quad x^T x = 1. \tag{20}$$

The TLS algorithm should be applied on the matrix $\mathcal{A}$ which has the structure (10) with

$$\alpha_k^T = \left(\left[Z^T(j\omega_k) S^T(j\omega_k)\right] \otimes I_{ny}\right) L. \tag{21}$$

The one-dimensional TLS estimates are inconsistent and suffer from the same problems as the multidimensional solution.

3.2. GTLS. Putting $W = I_m$ in (13) gives the *multidimensional* GTLS estimates [9]

$$V_{GTLS}(X,Z) = \sum_{k=1}^{F} \text{trace}\left(\varepsilon^H(j\omega_k,X)\left[\sum_{l=1}^{F} W_{ML}^{-2}(j\omega_l,X)\right]^{-1}\varepsilon(j\omega_k,X)\right) \tag{22}$$

subject to $X^TX = I_d$.

Calculating the expected value of (22) gives

$$E\{V_{GTLS}(X,Z)\} = V_{GTLS}(X,Z_0) + ny. \tag{23}$$

Since the expected value of the cost function is minimal in the true parameter values ($V_{GTLS}(X,Z_0) = 0$), the estimates are strongly consistent under Assumption 2.1 (provided that the regularity assumptions of section 5.2 are satisfied). Due to the equal weighting $\sum_{l=1}^{F} W_{ML}^{-2}(j\omega_l,X)$ of the residuals $\varepsilon(j\omega_k,X)$ over all frequencies in (22), the GTLS estimates overemphasize the high frequency errors [11], [14]. It explains why its efficiency can be very poor.

An analytic expression B for the square root of the column covariance matrix of A can be found [21]

$$B = [\text{real}(\beta_1), \text{real}(\beta_2), \ldots, \text{real}(\beta_F), \text{imag}(\beta_1), \ldots, \text{imag}(\beta_F)]^T, \tag{24}$$

with

$$\beta_k^T = C_Z(j\omega_k)^{T/2} S^T(j\omega_k). \tag{25}$$

$C_Z(j\omega_k)^{1/2}$ is a square root of $C_Z(j\omega_k)$ and can be calculated by means of a Cholesky or singular value decomposition. The number of rows of the rectangular matrix B ($2F(nu+ny)$) can be quite large compared with the dimension of C ($ny(od+1)+nu(on+1)$). This is the reason why the GSVD of the matrix pair (A,C), with C a square root of B^TB, is usually preferred over that of (A,B).

Putting $W = I_{md}$ in (15) gives the *one-dimensional* GTLS estimates

$$V_{GTLS}(x,Z) = \frac{\sum_{k=1}^{F} \text{trace}(\varepsilon^H(j\omega_k,X)\varepsilon(j\omega_k,X))}{\sum_{l=1}^{F} \text{trace}(W_{ML}^{-2}(j\omega_l,X))} \tag{26}$$

$$\text{subject to } \text{vec}(X^T) = Lx \quad \text{and} \quad x^Tx = 1.$$

x_{GTLS} is calculated using the matrix pair $(\mathcal{A},C)$ or $(\mathcal{A},B)$ with $\mathcal{A}$ defined by (21) and where B has structure (24) with

$$\beta_k^T = ([C_Z(j\omega_k)^{T/2} S^T(j\omega_k)] \otimes I_{ny})L. \tag{27}$$

Calculating the expected value of (26) gives

$$E\{V_{GTLS}(x,Z)\} = V_{GTLS}(x,Z_0) + 1. \tag{28}$$

Since (28) is minimal in the true parameter values ($V_{GTLS}(x,Z_0) = 0$), the estimates are strongly consistent under Assumption 2.1 (provided that the regularity assumptions of section 5.2 are satisfied).

3.3. WGTLS. It is possible to introduce frequency-dependent weights in the *multidimensional* GTLS by multiplying each row of A with a frequency-dependent, real-valued weighting function $w_\varepsilon(j\omega_k)$. This results in the following W-matrix

$$W = I_2 \otimes \text{block diag}(I_M w_\varepsilon(j\omega_1), I_M w_\varepsilon(j\omega_2), \ldots, I_M w_\varepsilon(j\omega_F)). \tag{29}$$

Putting W (29) in (13) gives

$$\begin{aligned} &V_{WGTLS}(X,Z) \\ &= \sum_{k=1}^{F} \text{trace}\left(w_\varepsilon^2(j\omega_k)\varepsilon^H(j\omega_k,X)\left[\sum_{l=1}^{F} w_\varepsilon^2(j\omega_l)W_{ML}^{-2}(j\omega_l,X)\right]^{-1}\varepsilon(j\omega_k,X)\right) \\ &\text{subject to } X^TX + I_d. \end{aligned} \tag{30}$$

Note that all the entries of the residual error matrix $\varepsilon(j\omega_k, X)$ in (30) are weighted with the same scalar frequency-dependent weighting function $w_\varepsilon(j\omega_k)$.

For the *one-dimensional* implementation it is possible to introduce a Hermitian-symmetric weighting matrix $W_\varepsilon(j\omega_k) \in \mathcal{C}^{ny\times ny}$ in the WGTLS cost function. Indeed, transforming the weighted equation error as follows [3],

$$\begin{aligned} &\text{vec}(W_\varepsilon(j\omega_k)X^TS(j\omega_k)Z(j\omega_k)) \\ &\quad = ([Z^T(j\omega_k)S^T(j\omega_k)] \otimes W_\varepsilon(j\omega_k))\text{vec}(X^T) \\ &\quad = (I_M \otimes W_\varepsilon(j\omega_k))([Z^T(j\omega_k)S^T(j\omega_k)] \otimes I_{ny})\text{vec}(X^T) \end{aligned} \tag{31}$$

leads to a one-dimensional WGTLS problem with weighting matrix

$$W = \begin{bmatrix} \text{real}(W_c) & -\text{imag}(W_c) \\ \text{imag}(W_c) & \text{real}(W_c) \end{bmatrix}, \tag{32}$$

with

$$W_c = \text{block diag}(I_M \otimes W_\varepsilon(j\omega_1), I_M \otimes W_\varepsilon(j\omega_2), \ldots, I_M \otimes W_\varepsilon(j\omega_F)). \tag{33}$$

W is symmetric since $W_c^H = W_c$. Putting these expressions in (15) gives, after some calculations,

$$\begin{aligned} &V_{WGTLS}(x,Z) = \frac{\sum_{k=1}^{F}\text{trace}(\varepsilon^H(j\omega_k,X)W_\varepsilon^2(j\omega_k)\varepsilon(j\omega_k,X))}{\sum_{l=1}^{F}\text{trace}(W_\varepsilon^2(j\omega_l)W_{ML}^{-2}(j\omega_l,X))} \\ &\text{subject to } \text{vec}(X^T) = Lx \text{ and } x^Tx = 1. \end{aligned} \tag{34}$$

x_{WGTLS} is calculated using the matrix pair $(\mathcal{A}_W, C)$ or $(\mathcal{A}_W, B)$, where the matrices $\mathcal{A}_W = W\mathcal{A}$ and B have, respectively, structure (10) and (24) with

$$\begin{aligned} &\alpha_k^T = ([Z^T(j\omega_k)S^T(j\omega_k)] \otimes W_\varepsilon(j\omega_k))L \\ &\quad \text{and} \quad \beta_k^T = ([C_Z(j\omega_k)^{T/2}S^T(j\omega_k)] \otimes W_\varepsilon(j\omega_k))L. \end{aligned} \tag{35}$$

3.4. BTLS. Adding an appropriate frequency-dependent weighting to the GTLS estimator is the key solution to improve its efficiency. The ML solution calculated under Assumptions 2.1 and 2.2 [9],

$$V_{ML}(X,Z) = \sum_{k=1}^{F} \operatorname{trace}(\varepsilon^H(j\omega_k,X) W_{ML}^2(j\omega_k,X)\varepsilon(j\omega_k,X)) \quad \text{subject to } X^TX = I_d, \tag{36}$$

learns that the optimal left weighting of the residual of the model $\varepsilon(j\omega_k, X)$ equation equals $W_{ML}(j\omega_k, X)$. X is unfortunately unknown so that only an approximation $W_{ML}(j\omega_k, \hat{X})$ of the optimal weighting can be calculated through an initial guess $\hat{X}$ of the model parameters.

Left multiplication of the residuals $\varepsilon(j\omega_k, X)$ (17) with $W_{ML}(j\omega_k, \hat{X})$ gives an expression $W_{ML}(j\omega_k, \hat{X})X^TS(j\omega_k)Z(j\omega_k)$ which can no longer be written under the form $AX = 0$. Hence it is impossible to apply the full ML weighting in the *multi-dimensional* WGTLS estimator. Only scalar functions of $W_{ML}(j\omega_k, \hat{X})$ are allowed in (30). Functions that work reasonably well are

$$w_\varepsilon(j\omega_k) = \sqrt[2\cdot ny]{\det(W_{ML}(j\omega_k,\hat{X}))} \quad \text{or} \quad w_\varepsilon(j\omega_k) = \sqrt{\operatorname{trace}(W_{ML}(j\omega_k,\hat{X}))}\,. \tag{37}$$

Putting expressions (37) in (30) defines the one-step multidimensional BTLS estimates.

The *one-dimensional* WGTLS solution allows us to include the full ML weighting. Replacing $W_\varepsilon(j\omega_k)$ by $W_{ML}(j\omega_k, \hat{X})$ in (34) gives the one-step, one-dimensional BTLS estimates [9]

$$V_{BTLS}(x,Z,\hat{X}) = \frac{\sum_{k=1}^{F} \operatorname{trace}(\varepsilon^H(j\omega_k,X) W_{ML}^2(j\omega_k,\hat{X})\varepsilon(j\omega_k,X))}{\sum_{l=1}^{F} \operatorname{trace}(W_{ML}^2(j\omega_l,\hat{X}) W_{ML}^{-2}(j\omega_l,X))} \quad \text{subject to } x^Tx = 1 \text{ and } \operatorname{vec}(X^T) = Lx. \tag{38}$$

Assuming that the initial guess $\hat{X}$ is independent of the measurements $Z(j\omega_k)$, $k = 1, 2, \ldots, F$, the expected value of (38) equals

$$E\{V_{BTLS}(x,Z,\hat{X})\} = V_{BTLS}(x,Z_0,\hat{X}) + 1. \tag{39}$$

Since (39) is minimal in the true model parameters ($V_{BTLS}(x, Z_0, \hat{X}) = 0$), the one-step BTLS estimator is strongly consistent (see section 5.2). Due to the appropriate frequency weighting, the estimates (38) have nearly ML efficiency [24], [14], [9]. The intuitive explanation for this is the close resemblence between the BTLS and ML cost functions: replacing $\hat{X}$ by X in (38) gives $V_{BTLS}(x, Z, X) = V_{ML}(X, Z)/(Fny)$. The efficiency can be improved on further by using $\hat{x} = x_{BTLS}$ to calculate an improved weighting, recalculating the BTLS estimates, and so on until convergence is obtained. As starting value for this iterative procedure, a strongly consistent estimate $\hat{x}$ is used; for example, $\hat{x} = x_{GTLS}$. Although $\hat{X}$ now clearly depends in each step on the measurements $Z(j\omega_k)$, $k = 1, 2, \ldots, F$, the resulting multistep BTLS estimate is still strongly consistent under Assumption 2.1 (provided that the regularity assumptions of section 5.2 are satisfied). Indeed, since $\hat{X}$ converges strongly to X_0, the cost function (38) converges strongly ($F \to \infty$) to $V_{BTLS}(x, Z_0, X_0) + 1$ which is minimal in the true model parameters x_0.

3.5. WTLS for scalar systems. A disadvantage of the BTLS estimator is that it is not self-starting: an initial guess of the model parameters should be available to calculate a "reasonable" weighting. The question now is: can a "reasonable" weighting be obtained without any prior knowledge of the model parameters? For scalar systems ($nu = ny = 1$), a solution has been found. The following weighting approximates the optimal ML weighting $W_{ML}(j\omega_k, x)$ in a nonparametric way [15]:

$$W_\varepsilon^{-2}(j\omega_k, G_k) = \left(\frac{\sigma_Y^2(j\omega_k)}{|G(j\omega_k)|} + |G(j\omega_k)|\sigma_U^2(j\omega_k) - 2\text{real}\left(\sigma_{YU}^2(j\omega_k)e^{-j\arg(G(j\omega_k))}\right)\right) T(\Omega_k), \tag{40}$$

with $G(j\omega_k) = Y(j\omega_k)/U(j\omega_k)$ the measured frequency response function and

$$T(\Omega_k) = \frac{(|\Omega_k|^{(on+1)} - 1)}{(|\Omega_k| - 1)} \frac{(|\Omega_k|^{(od+1)} - 1)}{(|\Omega_k| - 1)}. \tag{41}$$

Note that $T(\Omega_k) = (on+1)(od+1)$ for discrete time systems (z-domain) and for each value $|\Omega_k| = 1$. Using (40), and defining the weighting matrix W as in (29), one can construct a WTLS estimator

$$V_{WTLS}(x, U, Y) = \sum_{k=1}^{F} W_\varepsilon^2(j\omega_k, G_k) \frac{|\varepsilon(j\omega_k, x)|^2}{x^T x} \quad \text{subject to } x^T x = 1, \tag{42}$$

and a WGTLS estimator

$$V_{WGTLS}(x, U, Y) = \frac{\sum_{k=1}^{F} W_\varepsilon^2(j\omega_k, G_k)|\varepsilon(j\omega_k, x)|^2}{\sum_{l=1}^{F} W_\varepsilon^2(j\omega_l, G_l) W_{ML}^{-2}(j\omega_l, x)} \quad \text{subject to } x^T x = 1. \tag{43}$$

Both estimates are inconsistent since the weighting $W_\varepsilon(j\omega_k, G_k)$ is a function of the measurement noise. Among the existing methods, the proposed estimators (42) and (43), considered as a pair, lead to better, or at least not worse, starting values for the BTLS algorithm (38) [15].

4. Extensions.

4.1. Identification from transfer function matrix measurements. Sometimes the input/output Fourier data are not available and the identification should start from measured transfer function matrices $G(j\omega_k)$ [17]. The model equation is given by (4) with $M = nu$, $Y_0(j\omega) = G_0(j\omega)$, and $U_0(j\omega) = I_{nu}$. The noise on the transfer function measurements $\Delta Z^T(j\omega_k) = [\Delta G^T(j\omega_k)\ 0]$ satisfies Assumption 2.1 so that all the multidimensional TLS estimators developed in section 3 can still be applied. This is no longer the case for the one-dimensional WGTLS implementations (except for the one-dimensional TLS) since, in general, the columns of $\Delta G(j\omega_k)$ are correlated and have different covariance matrices. Therefore the one-dimensional implementations need more noise information in case of transfer function matrix measurements.

ASSUMPTION 4.1. *$\Delta g(j\omega_k) = \text{vec}(\Delta G(j\omega_k)), k = 1, \ldots, F$ are zero-mean, mixing (over the frequency), complex distributed random vectors with known Hermitian-symmetric matrices $E\{\Delta g(j\omega_k)\Delta g^H(j\omega_k)] = C_{\text{vec}(G)}(j\omega_k)$, satisfying $E\{\Delta g(j\omega_k)\ \Delta g^T(j\omega_k)\} = 0$.*

ASSUMPTION 4.2. *The errors* $\Delta g(j\omega_k) = \mathrm{vec}(\Delta G(j\omega_k))$ *are independent (over the frequency) complex normally distributed random vectors.*

Under Assumptions 4.1 and 4.2, the ML solution becomes

$$V_{ML}(X,Z) = \sum_{k=1}^{F} \mathrm{vec}^H(\varepsilon(j\omega_k, X)) W_{ML}^2(j\omega_k, X) \mathrm{vec}(\varepsilon(j\omega_k, X)) \quad \text{subject to } X^T X = I_d, \tag{44}$$

where

$$\begin{aligned} W_{ML}^{-2}(j\omega_k, X) &= E\{\mathrm{vec}(\Delta\varepsilon(j\omega_k, X))\mathrm{vec}^H(\Delta\varepsilon(j\omega_k, X))\} \\ &= (I_{nu} \otimes D(\Omega_k, X)) C_{\mathrm{vec}(G)}(j\omega_k)(I_{nu} \otimes D^H(\Omega_k, X)). \end{aligned} \tag{45}$$

For the one-dimensional WGTLS solution we use a left weighting matrix W which has structure (32) with $W_c = \text{block diag}(W_\varepsilon(j\omega_1), W_\varepsilon(j\omega_2), \ldots, W_\varepsilon(j\omega_F))$ and $W_\varepsilon(j\omega) \in \mathcal{C}^{(nu\cdot ny)\times(nu\cdot ny)}$ a Hermitian-symmetric matrix. Putting these expressions in (15) gives

$$V_{WGTLS}(x,Z) = \frac{\sum_{k=1}^{F} \mathrm{vec}^H(\varepsilon(j\omega_k, X)) W_\varepsilon^2(j\omega_k) \mathrm{vec}(\varepsilon(j\omega_k, X))}{\sum_{l=1}^{F} \mathrm{trace}(W_\varepsilon^2(j\omega_l) W_{ML}^{-2}(j\omega_l, X))} \quad \text{subject to } \mathrm{vec}(X^T) = Lx \text{ and } x^T x = 1. \tag{46}$$

Replacing $W_\varepsilon^2(j\omega_k)$ by $W_{ML}(j\omega_k, \hat{X})$ in (46) gives the BTLS estimator

$$V_{BTLS}(x,Z,\hat{X}) = \frac{\sum_{k=1}^{F} \mathrm{vec}^H(\varepsilon(j\omega_k, X)) W_{ML}^2(j\omega_k, \hat{X}) \mathrm{vec}(\varepsilon(j\omega_k, X))}{\sum_{l=1}^{F} \mathrm{trace}(W_{ML}^2(j\omega_l, \hat{X}) W_{ML}^{-2}(j\omega_l, X))} \quad \text{subject to } \mathrm{vec}(X^T) = Lx \text{ and } x^T x = 1. \tag{47}$$

Under Assumption 4.1, the properties of the one-dimensional WGTLS (46) and BTLS (47) estimates are the same as those of section 3. The special case of independently measured entries $g_{ij}(j\omega_k)$ of the transfer function matrix $G(j\omega_k)$ is handled in [17].

4.2. High order systems. Transfer function model (3) leads to an ill-conditioned matrix WA for model orders $od \geq 40$. The numerical conditioning of WA can significantly be improved by expanding the numerator and denominator of (3) in orthogonal Forsythe polynomial matrices [7], [22]. Using this approach, very high order scalar systems $od \geq 120$ have been identified on experimental data [16].

4.3. Complex systems. The results of section 3 can be generalized to transfer function models with complex coefficients ($N_k \in \mathcal{C}^{ny\times nu}$, $D_k \in \mathcal{C}^{ny\times ny}$). Therefore it is sufficient to write model equation (4) at the excited DFT frequencies as a set of F complex equations

$$A_0 X = 0 \quad (A_0 \in \mathcal{C}^{m\times n}, X \in \mathcal{C}^{n\times d}, m = F, n = (od{+}1)ny{+}(on{+}1)nu, d = ny), \tag{48}$$

with $A_0 = [\alpha_{01}, \alpha_{02}, \ldots, \alpha_{0F}]^T$ (see (7) for the definition of α_{0k}) and to replace the transpose operator T at the appropriate places in section 3 by the Hermitian transpose operator H. For the WGTLS and BTLS estimators, the real-valued left weighting

W (32) is replaced by the complex-valued weighting W_c (33). Making the changes $X^TX \to X^HX$ and $x^Tx \to x^Hx$, the expressions for the cost functions (20), (26), (34), (36), (38), and (43) remain valid. A potential application of rational functions with complex coefficients (scalar case) is the modeling of nuclear magnetic resonance spectra.

5. The WGTLS estimator using the sample noise covariance matrix.

5.1. Introduction.

The WGTLS solution

$$\arg\min_{\hat{A},X} \|W(A-\hat{A})C^{-1}\|_F^2 \quad \text{subject to } \hat{A}X = 0 \quad \text{and} \quad X^TX = I_d, \tag{49}$$

with $X \in \mathcal{R}^{n\times d}$ produces consistent estimates if C satisfies $C^TC = E\{\Delta A^TW^TW\Delta A\}$. The covariance matrix C^TC is in most practical applications unknown. This section describes the asymptotic properties of the WGTLS estimates when the true noise covariance matrix C^TC is replaced by the sample noise covariance matrix obtained from a small number $K > 1$ of independent realizations A of the true unknown matrix A_0.

The main advantages of this approach are its robustness w.r.t. incorrect noise assumptions and the fact that no parametric noise model should be estimated (no noise model order selection and no parametric noise model should even exist). The only price to pay is that the measurements need to be repeated at least two times.

In order to simplify the asymptotic analysis ($m \to \infty$) of the WGTLS estimator using the sample noise covariance matrix (SWGTLS), the main steps in the analysis of the WGTLS estimator using the exact noise covariance matrix are first given.

5.2. WGTLS estimator—exact noise covariance matrix.

The analysis of the WGTLS estimator using the exact noise covariance matrix fits within the general framework of [23]. Therefore only the basic assumptions and the main results will be given here. The cost function interpretation of the WGTLS says that when X^TC^TCX is regular, then

$$\arg\min_X \ \operatorname{trace}((X^TC^TCX)^{-1}(X^TA^TW^TWAX)) \tag{50}$$

is equivalent to the WGTLS solution. To analyze the asymptotic behavior of the estimates, one needs to make the following assumptions.

ASSUMPTION 5.1. *For all $m > n$, the entries of A are jointly mixing (over m) of order 4, W is deterministic with $\|W\|_1 < \infty$, X^TC^TCX is regular, and C is deterministic with $\|C\|_1 < \infty$.*

ASSUMPTION 5.2. *A_0 and the zero-mean perturbation ΔA are mutually independent and $C^TC = E\{\Delta A^TW^TW\Delta A\}$.*

ASSUMPTION 5.3. *There exists an exact model $A_0X_0 = 0$.*

ASSUMPTION 5.4. *The excitation is persistent: $\frac{1}{m}A_0^TW^TWA_0$ is of rank $n-d$ for any $m > n$ ($m = \infty$ included).*

THEOREM 5.1. *Under Assumptions* 5.1, 5.2, 5.3, *and* 5.4

$$\arg\min_X \operatorname{trace}((X^TC^TCX)^{-1}(X^TA^TW^TWAX))$$

is a strongly consistent estimate.

Proof. Uniform (w.r.t. X) convergence with probability one of the WGTLS cost toward its expected value is guaranteed under Assumption 5.1 [23]. Under Assump-

tion 5.2, the expected value of the GTLS cost (50) becomes

$$
\begin{aligned}
&E\{\operatorname{trace}((X^TC^TCX)^{-1}(X^TA^TW^TWAX))\}\\
&= \operatorname{trace}((X^TC^TCX)^{-1}(X^TA_0^TW^TWA_0X)) + d.
\end{aligned}
\tag{51}
$$

The cost (51) is minimal in the exact parameters X_0 under Assumption 5.3. Assumption 5.4 guarantees that all the global minima of (51) satisfy $A_0X = 0$. Strong consistency of the estimates X_{WGTLS} immediately follows[3] [23]: a.s. $\lim_{m\to\infty} X_{WGTLS} = X_0$. ∎

ASSUMPTION 5.5. *For all $m > n$, the entries of A are jointly mixing (over m) of order ∞.*

THEOREM 5.2. *Under Assumptions* 5.1–5.5, $\arg\min_X \operatorname{trace}((X^TC^TCX)^{-1}(X^TA^TW^TWAX))$ *converges in law ($m \to \infty$) to a Gaussian random variable with mean value X_0.*

Proof. For the proof, see [23]. ∎

Assumption 5.5 is, for example, satisfied for independent identical distributed noise of the exponential family distribution [20] passing through a linear stable filter.

5.3. SWGTLS estimator—sample noise covariance matrix. The analysis requires an additional assumption and will be done in two steps.

ASSUMPTION 5.6. *A strongly consistent estimate D, independent of X, of the noise covariance matrix C^TC is available:* a.s. $\lim_{m\to\infty}(D - C^TC)/m = 0$.

First the WGTLS estimator will be studied, assuming that a strongly consistent estimate D of the exact noise covariance matrix is available. Next it will be shown that the sample covariance matrix of the noise obtained from independent repeated experiments satisfies Assumption 5.6.

THEOREM 5.3. *Under Assumptions* 5.1–5.4 *and* 5.6, $\arg\min_X \operatorname{trace}((X^TDX)^{-1}(X^TA^TW^TWAX))$ *is a strongly consistent estimate (for $m \to \infty$).*

Proof. Following along the lines of section 5.2, strong consistency of the estimates is proven if it can be shown that the WGTLS cost function using the estimated noise covariance matrix D converges uniformly with probability one to an expression which is minimal in the true parameter values. Under Assumptions 5.1, 5.2, and 5.6 the WGTLS cost function using the estimated noise covariance matrix D converges uniformly with probability one to the expected value of the cost function using the exact noise covariance matrix (51),

$$
\begin{aligned}
&\text{a.s.} \lim_{m\to\infty} \operatorname{trace}((X^TDX)^{-1}(X^TA^TW^TWAX)\\
&-E\{(X^TC^TCX)^{-1}(X^TA^TW^TWAX)\})\\
&= \operatorname{trace}\left(\left(X^T\left(\text{a.s.}\lim_{m\to\infty}\frac{1}{m}D\right)X\right)^{-1}X^T\left(\text{a.s.}\lim_{m\to\infty}\frac{1}{m}(A^TW^TWA)\right)X\right)\\
&-\operatorname{trace}\left(\left(X^T\left(\lim_{m\to\infty}\frac{1}{m}C^TC\right)X\right)^{-1}X^T\left(\lim_{m\to\infty}\frac{1}{m}(A_0^TW^TWA_0)\right)X\right) - d\\
&= 0.
\end{aligned}
\tag{52}
$$

[3] a.s. lim stands for "almost sure limit" or "limit with probability one"; it means that the event "$\lim_{m\to\infty}$ converges" happens with probability one. See [19] for a formal definition.

Combining this result with Assumptions 5.3 and 5.4 proves the strong consistency. ∎

Assume now that repeated observations of the true unknown matrix A_0 are available, $A^{[k]} = A_0 + \Delta A^{[k]}$ $(k = 1, \ldots, K)$, and use the mean value $\bar{A} = \sum_{k=1}^{K} A^{[k]}/K$ for the identification (put $A = \bar{A}$ in (50)).

ASSUMPTION 5.7. *For all $m > n$ and for all k, the entries of $A^{[k]}$ are jointly mixing (over m) of order 4 and W is deterministic with $\|W\|_1 < \infty$.*

ASSUMPTION 5.8. *A_0 is a constant (k-independent) matrix which is stochastically independent of $\Delta A^{[k]}$ for all k. $\Delta A^{[k]}$ is the kth independent realization $(k = 1, 2, \ldots, K)$ of a zero-mean, noise process with k-independent covariance matrix.*

THEOREM 5.4. *Under Assumptions* 5.7 *and* 5.8, $D = \frac{1}{K(K-1)} \sum_{k=1}^{K} (A^{[k]} - \bar{A})^T W^T W (A^{[k]} - \bar{A})$ *is a strongly consistent estimate (for $m \to \infty$) of the column covariance matrix $C^T C$ of $\bar{A}$.*

Proof. Define $B^{[k]} = WA^{[k]}$ with matrix elements $b_{ij}^{[k]}$. The ijth element of D is then given by

$$d_{ij} = \frac{m}{K(K-1)} \sum_{k=1}^{K} \frac{1}{m} \sum_{p=1}^{m} \left(b_{pi}^{[k]} - \overline{b_{pi}} \right) \left(b_{pj}^{[k]} - \overline{b_{pj}} \right) \quad \text{with} \quad \overline{b_{ij}} = \frac{1}{K} \sum_{k=1}^{K} b_{ij}^{[k]}. \tag{53}$$

Assumption 5.7 guarantees that the individual terms $(b_{pi}^{[k]} - \overline{b_{pi}})(b_{pj}^{[k]} - \overline{b_{pj}})$ of (53) are jointly mixing (over m) of order 2 [5]. Almost sure convergence of $\sum_{p=1}^{m} (b_{pi}^{[k]} - \overline{b_{pi}})(b_{pj}^{[k]} - \overline{b_{pj}})/m$, and hence also of d_{ij}/m, toward its expected value then follows by applying the strong law of large numbers for mixing sequences (see Lemma 3 of [23]). Assumption 5.8 guarantees that the sample covariance is an unbiased estimate for finite K, and hence a.s. $\lim_{m\to\infty}(D - E\{D\})/m =$ a.s. $\lim_{m\to\infty}(D - C^T C)/m = 0$. ∎

ASSUMPTION 5.9. *$X^T C^T C X$ is regular with $\|C\|_1 < \infty$.*

THEOREM 5.5. *Under Assumptions* 5.3, 5.4, *and* 5.7–5.9,

$$\arg\min_X \operatorname{trace}((X^T D X)^{-1} (X^T \bar{A}^T W^T W \bar{A} X)),$$

with D given by (53), *is a strongly consistent estimate ($m \to \infty$).*

Proof. Apply Theorems 5.3 and 5.4. ∎

ASSUMPTION 5.10. *For all $m > n$ and for all k, the entries of $A^{[k]}$ are jointly mixing (over m) of order ∞.*

THEOREM 5.6. *Under Assumptions* 5.3, 5.4, *and* 5.7–5.10,

$$\arg\min_X \operatorname{trace}((X^T D X)^{-1} (X^T \bar{A}^T W^T W \bar{A} X))$$

converges in law ($m \to \infty$) to the same asymptotic Gaussian distribution as the WGTLS estimates using the exact covariance matrix.

Proof. Apply Theorem 6 of [23]. ∎

Intuitively, this result is motivated by observing that the estimates are normally distributed for any deterministic value of $C^T C$. Since D given by (53) converges sufficiently fast to $E\{\Delta \bar{A}^T W^T W \Delta \bar{A}\}$, the stochastic variation in the estimate D will become negligible w.r.t. the Gaussian stochastic variation of the estimates when using the exact value of $E\{\Delta \bar{A}^T W^T W \Delta \bar{A}\}$.

Since (53) exists for any $K \geq 2$, it follows that the asymptotic properties (strong consistency, asymptotic covariance matrix, and asymptotic normality) of the GTLS

estimator using the sample covariance matrix apply for any $K \geq 2$. Hence only two repeated independent experiments are sufficient to solve the errors-in-variables problem within a TLS framework.

6. The BTLS estimator using the sample noise covariance matrix. The reasoning will be held for the one-dimensional BTLS estimates (38) using measured input/output Fourier data. Extension to the transfer function measurement case (47) is straightforward.

Following along the lines of section 5.3, one could think to replace the true noise covariance matrix $C_Z(j\omega)$ everywhere in (38) by the sample noise covariance matrix

$$\hat{C}_Z(j\omega) = \frac{1}{K(K-1)} \sum_{k=1}^{K} (Z^{[k]}(j\omega) - \bar{Z}(j\omega))(Z^{[k]}(j\omega) - \bar{Z}(j\omega))^H, \tag{54}$$

where $\bar{Z}(j\omega) = \sum_{k=1}^{K} Z^{[k]}(j\omega)/K$ is the sample mean ($Z^{[k]}(j\omega)$, $\bar{Z}(j\omega) \in \mathcal{C}^{(ny+nu)\times M}$). Proceeding in that way we violate the assumptions of the framework developed in [23]. Indeed, strong consistency is only guaranteed by Theorem 4 of [23] if the number of stochastic parameters in the weighting remains finite for finite K and $F \to \infty$ and if they converge strongly ($F \to \infty$) to some limit value which is independent of the estimated model parameters $\hat{x}$. Therefore, to preserve the strong consistency, the noise covariance matrix $C_Z(j\omega)$ in the left weighting matrix W should be modeled over the frequency using a finite (F-independent) number of parameters θ. The estimates $\hat{\theta}$ should strongly converge to some limit value θ_*, independent of the model parameters x. As it is the case for the GTLS estimator (see section 5.3) the right weighting matrix C must still be calculated using the original sample covariance matrices (54).

For computational reasons, only noise models $C_Z(j\omega, \theta)$ which are linear in the parameters are considered. For example, for the r, sth entry we get

$$(C_Z(j\omega, \theta))_{rs} = \sum_{k=1}^{p_{rs}} \theta_k^{rs} h_k^{rs}(j\omega) \qquad r, s = 1, 2, \ldots, nu + ny, \tag{55}$$

where $h_k^{rs}(j\omega)$, $k = 1, 2, \ldots, p_{rs}$ are linear independent basis functions and with p_{rs} independent of F. Under Assumptions 5.7 and 5.8 the linear least squares estimate $\hat{\theta}^{rs}$ of the vector of the noise model parameters θ^{rs},

$$\begin{aligned} &\hat{\theta}^{rs} = (H^{rsT} H^{rs})^{-1} H^{rsT} [(\hat{C}_Z(j\omega_1))_{rs}, \ldots, (\hat{C}_Z(j\omega_F))_{rs}]^T \\ &\text{with } (H^{rs})_{kl} = h_k^{rs}(j\omega_l), \end{aligned} \tag{56}$$

converges strongly ($F \to \infty$) to some x-independent limit value θ_*^{rs}. The estimated noise model for entry r, s,

$$[(C_Z(j\omega_1, \hat{\theta}))_{rs}, \ldots, (C_Z(j\omega_F, \hat{\theta}))_{rs}]^T = H^{rs} \hat{\theta}^{rs}, \tag{57}$$

represents a linear projection of an F-dimensional space onto a p_{rs}-dimensional space.

Replacing in (38) $C_Z(j\omega)$ by $C_Z(j\omega, \hat{\theta})$ in the left weighting W, $C_Z(j\omega)$ by $\hat{C}_Z(j\omega)$ in the right weighting C, and $Z(j\omega)$ by the sample mean $\bar{Z}(j\omega)$ defines the one-dimensional, one-step BTLS using the sample noise covariance matrix (SBTLS)

$$V_{SBTLS}(x, \bar{Z}, \hat{X}, \hat{\theta}) = \frac{\sum_{k=1}^{F} \operatorname{trace}(\varepsilon^H(j\omega_k, X) W_{PML}^2(j\omega_k, \hat{X}, \hat{\theta}) \varepsilon(j\omega_k, X))}{\sum_{l=1}^{F} \operatorname{trace}(W_{PML}^2(j\omega_l, \hat{X}, \hat{\theta}) W_{SML}^{-2}(j\omega_l, X))} \tag{58}$$

subject to $x^T x = 1$ and $\operatorname{vec}(X^T) = Lx$.

TABLE 1
Coefficients of the transfer function of the fifth-order butterworth filter with a transmission zero.

n_0	n_1	n_2	d_0	d_1
1	0	1/9	1	0.449941
d_2	d_3	d_4	d_5	d_2
0.101223	1.40740e-2	1.20939e-3	5.19623e-5	0.101223

$W_{PML}(j\omega, \hat{X}, \hat{\theta})$ and $W_{SML}(j\omega, X)$ stand for the ML weighting (19) calculated with, respectively, the parametric noise model $C_Z(j\omega, \hat{\theta})$ and the sample covariance matrix $\hat{C}_Z(j\omega)$,

$$\begin{aligned} W_{PML}(j\omega, \hat{X}, \hat{\theta}) &= \hat{X}^T S(j\omega_k) C_Z(j\omega, \hat{\theta}) S^H(j\omega_k) \hat{X}, \\ W_{SML}(j\omega, X) &= X^T S(j\omega_k) \hat{C}_Z(j\omega) S^H(j\omega_k) X. \end{aligned} \tag{59}$$

Under the assumptions of section 5.3, the one-step and multistep (see section 3.4) SBTLS cost functions (58) converge strongly ($F \to \infty$ and $K > 1$) to, respectively, $V_{SBTLS}(x, Z_0, \hat{X}, \theta_*)$ and $V_{SBTLS}(x, Z_0, X_0, \theta_*)$, which are minimal in x_0. The one-step and multistep SBTLS estimates are hence strongly consistent. The efficiency of the SBTLS estimates strongly depends on the parametric noise model (56) used: the better the parametric model explains the noise (co)variances, the closer the SBTLS efficiency will approach the ML efficiency.

x_{SBTLS} is calculated using the matrix pair $(\mathcal{A}_W, C)$ or $(\mathcal{A}_W, B)$, where the matrices $\mathcal{A}_W = W\mathcal{A}$ and B have, respectively, structure (10) and (24) with

$$\begin{aligned} \alpha_k^T &= ([\bar{Z}^T(j\omega_k) S^T(j\omega_k)] \otimes W_{PML}(j\omega_k, \hat{X}, \hat{\theta}))L, \\ \beta_k^T &= ([\hat{C}_Z(j\omega_k)^{T/2} S^T(j\omega_k)] \otimes W_{PML}(j\omega_k, \hat{X}, \hat{\theta}))L. \end{aligned} \tag{60}$$

7. Simulation examples.

7.1. Single input, single output systems. Three simulation examples are shown in this section. The first illustrates the TLS, WTLS, GTLS, WGTLS, and multistep BTLS estimators with known noise covariance matrix, while the second and third compare the generalized and bootstrapped total least squares estimates using the sample noise covariance matrix (respectively, SGTLS and SBTLS) to those using the exact noise covariance matrix (respectively, GTLS and BTLS). For the three examples, the simulated plant is a fifth-order continuous time Butterworth filter with an extra transmission zero at $3/(2\pi)$ Hz. The coefficients of the transfer function are given in Table 1, and the amplitude and phase characteristics are shown in Figure 1.

In the first simulation, a data set of 100 equally distributed frequencies is generated in the band [0.05 Hz, 5 Hz] : $U_0(j\omega_k) = 1$, $Y_0(j\omega_k) = G(j\omega_k)$, $k = 1, 2, \ldots, 100$. Independent, zero-mean Gaussian noise with variance 2E–6 is added to the input and output spectra. One hundred disturbed data sets are generated. For each set the model parameters are calculated using the TLS (16), WTLS (42), GTLS (22), WGTLS (43), and BTLS (38) estimators under the constraint $n_1 = 0$ (the zero is forced to lie on the $j\omega$ axis), and $\|x\|_2 = 1$. For each set, the normalized squared residuals of the mean parameter estimates are calculated

$$(\bar{x} - x_0)^T C_x^{\dagger} (\bar{x} - x_0), \tag{61}$$

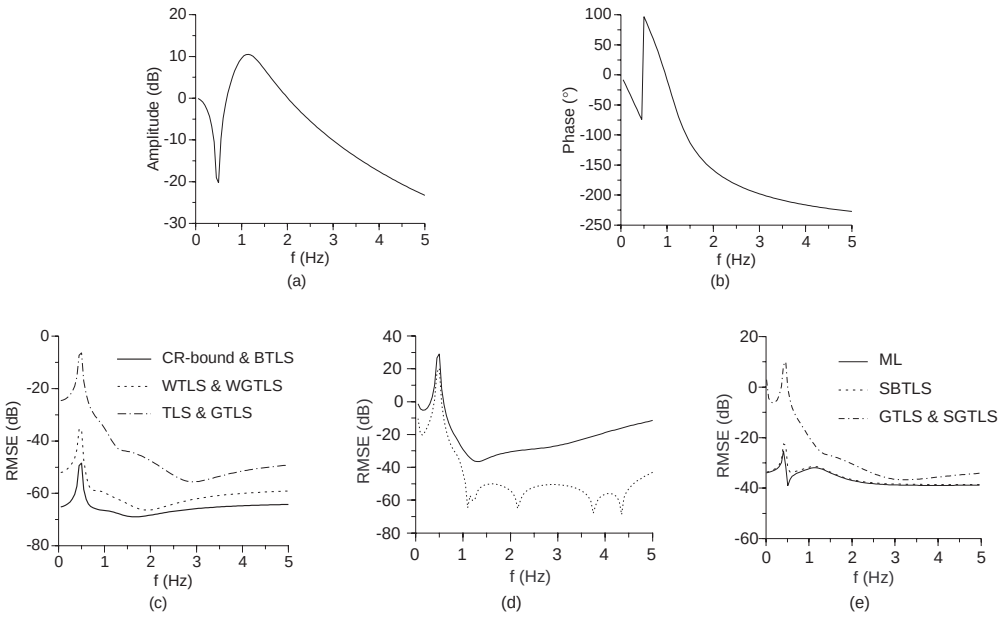


FIG. 1. *Fifth-order Butterworth filter with a transmission zero:* (a) *amplitude and* (b) *phase characteristics. Comparison of the relative mean square error (RMSE) of the transfer function estimates,* (c) *results simulation* 1, (d) *results simulation* 2: *GTLS estimates (solid line) and difference between the GTLS and SGTLS estimates (dashed line),* (e) *results simulation* 3.

with $\bar{x}$ the sample mean and C_x the sample covariance matrix of the model parameters x. If $\bar{x}$ is an unbiased Gaussian estimate of x_0 and if C_x equals the true covariance matrix, then (61) is (χ^2-distributed with a number of degrees of freedom equal to the number of parameters minus the number of constraints ($= 9-2 = 7$ in this simulation example). For 100 realizations of the model parameters, the sample mean will be fairly well normal distributed and the uncertainty on C_x is about 10%. This allows a bias test to be performed on the parameter estimates with a given confidence level. For example, the 95 percentile of a χ^2-random variable with seven degrees of freedom equals 14. According to Table 2, the GTLS and BTLS estimates are unbiased within a confidence level of 95%, while the TLS and WTLS estimates are biased within this confidence level. Although the WGTLS (43) estimates are inconsistent, no significant bias could be detected in the simulation (see Table 2). This is due to the high signal-to-noise ratio on the noisy frequency response function $G(j\omega)$. Using each set of 100 estimates of the model parameters, one can also calculate the relative mean square error of the transfer function estimate

$$\text{RMSE}(G) = E\{|(G - G_0)/G_0|^2)\}, \tag{62}$$

within an error of 1dB, and compare it to the Cramér–Rao lower bound on the relative transfer function error $(G - G_0)/G_0$. The results are shown in Figure 1c. The BTLS estimates coincide with the Cramér–Rao lower bound [6]. The large mean square error (MSE) of the TLS estimator is due to its bias (see Table 2), while that of the GTLS estimates are due to the variance (see Table 2). The same observation can be made for the pair WTLS and WGTLS (see Table 2 and Figure 1c).

In simulation 2, only the noise levels differ from simulation 1. The input noise, $\Delta U(j\omega)$, is chosen to be zero-mean Gaussian noise with variance equal to 1E-2. The

TABLE 2
Bias test on the model parameters—simulation 1.

estimator	eq. (61)	result bias test
TLS	2.2e3	biased
WTLS	16	biased
GTLS	6.4	unbiased
WGTLS	3.6	unbiased
BTLS	1.7	unbiased

output noise, $\Delta Y(j\omega)$, is a zero-mean Gaussian noise source with a variance of 1E-2 passing through a fourth-order Butterworth low-pass filter with its 3dB point at 1 Hz. The number of repeated, independent measurements equals two ($K = 2$). Figure 1d compares the generalized total least squares estimates using the sample noise covariance matrix estimated as described in section 5.3 (SGTLS) to those using the true noise covariance matrix (GTLS). It can be seen that both RMSE errors almost coincide. Applying the bias test (61) to both simulation results reveals that both estimates are unbiased within a confidence level of 95%. It confirms that two repeated, independent experiments are enough to replace the true noise covariance matrix in GTLS estimators with the noise sample covariance matrix, while maintaining the asymptotic properties.

In simulation 3, the noise levels and the number of frequencies differ from simulation 1: $\sigma_U^2(j\omega_k) = 4\text{E}-2$, $\sigma_Y^2(j\omega_k) = 4\text{E}-2|G_0(j\omega_k)|^2$, and $F = 500$. The ML, GTLS, SGTLS, multistep BTLS, and multistep SBTLS are calculated starting from two repeated independent experiments ($K = 2$). For the SBTLS estimates, the input and output variances (55) in the left weighting are modeled using 50 Gaussian basis functions $h_k^{rs}(j\omega) = \exp(-(\omega - \mu_k)^2/\zeta_k)$, with μ_k equispaced in the band [0.05 Hz, 5 Hz], $\zeta_k = 0.05, r, s = 1, 2$, and $k = 1, 2, \ldots, 50$. Figure 1e shows the results: the GTLS and SGTLS estimates coincide and the SBTLS estimates have almost ML efficiency. Applying the bias test (61) to all the simulation results reveals that all the estimates are unbiased within a confidence level of 95%.

7.2. MIMO systems. In this section, one simulation example is given to illustrate the difference between the multidimensional and the one-dimensional implementations that were proposed in section 3. This will be verified for the GTLS (a multidimensional implementation of the BTLS estimator is not available). Synthetic data is generated for a system with two inputs and three outputs. Independent, zero-mean, complex normally distributed noise is added to the synthetic input and output Fourier coefficients. The standard deviation of the noise equals 0.01 for the inputs and 0.1 for the outputs. The frequency band is [1 Hz, 100 Hz] and contains 100 equally distributed frequencies. Three estimations are considered: the multidimensional implementation of the GTLS (case 1), the one-dimensional GTLS with a monic denominator matrix polynomial (case 2), and the one-dimensional GTLS with the zero-order matrix coefficient of the denominator fixed to the identity matrix (case 3). In Figure 2a, the difference in dB between the true and the estimated transfer functions of entry (2, 2) is shown. The difference between the estimates are given in Figure 2b. From Figures 2a and 2b, one can conclude that, for the one-dimensional case, different constraints do indeed result in different estimates and that the one-dimensional results differ from the multidimensional one. The differences are,

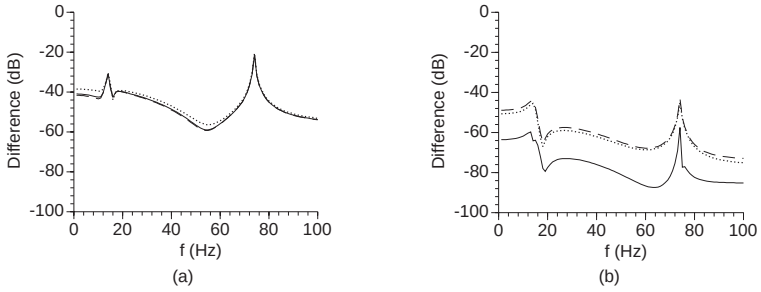


FIG. 2. *Influence of the parameter constraints.* (a) *Difference in dB between the true and estimated transfer function corresponding with entry* $(2,2)$*: case* 1 *(solid line), case* 2 *(dashed line), and case* 3 *(dotted line).* (b) *Difference in dB between the estimates of: cases* 1 *and* 2 *(solid line), cases* 2 *and* 3 *(dashed line), and cases* 1 *and* 3 *(dotted line).*

however, quite small.

8. Measurement example. Figure 3 shows measurements of the vibrations of the wings of an airplane (flight flutter data analysis) in the frequency band [4 Hz, 11 Hz]. The test was performed using a series of three short duration burst swept-sine [4 Hz, 40 Hz] excitations, sampled at 100 Hz. The data was measured using two channels corresponding to the force and the acceleration response, respectively. The sample mean and sample (co)variances of the three independent realizations ($K = 3$) of the input/output Fourier data are calculated in the band [4 Hz, 11 Hz]. The measurements have been modeled with a rational form $on = 11$, $od = 10$ using the SGTLS and multistep SBTLS estimators (see Figure 3). For SBTLS the (co)variances of the input/output DFT spectra in the left weighting are modeled by a constant: put $h_k^{rs}(j\omega) = 1$, $r, s = 1, 2$, and $k = 1$ in (55). The ML estimates have been added for comparison purposes. Since three independent realizations are not sufficient to use sample (co)variances within an ML framework [18], the noise covariance matrix for the ML estimate was obtained by analyzing the disturbing noise during the dead time in between consecutive bursts. From Figure 3, it follows that the SBTLS estimates have ML quality. The SGTLS estimates miss the second resonance peak which can be explained by the low signal-to-noise ratio of the measurements (about 10 dB on the transfer function) and its inappropriate weighting (see section 3.2). This measurement example nicely illustrates that a good frequency-dependent weighting of the residuals of the equation error (17) is of crucial importance to obtain good estimates.

9. Conclusions. The presented analysis of TLS estimators for frequency domain identification of multivariable systems leads to the following two main messages:

- Use the equivalent cost function minimized by the estimator to predict its asymptotic properties. Comparison of this cost with the maximum likelihood solution allows us to propose "optimal" left weighting matrices which significantly reduce the uncertainty of the TLS estimates.
- Two independent experiments with the same excitation signals are enough to replace the true noise covariance matrix in WGTLS and BTLS estimators by the noise sample covariance matrix while maintaining the asymptotic properties. A TLS method has been constructed (SBTLS) which does not require the prior knowedge of the noise covariance matrix and which has almost ML efficiency. Note that these design rules can also be applied to other identification problems.

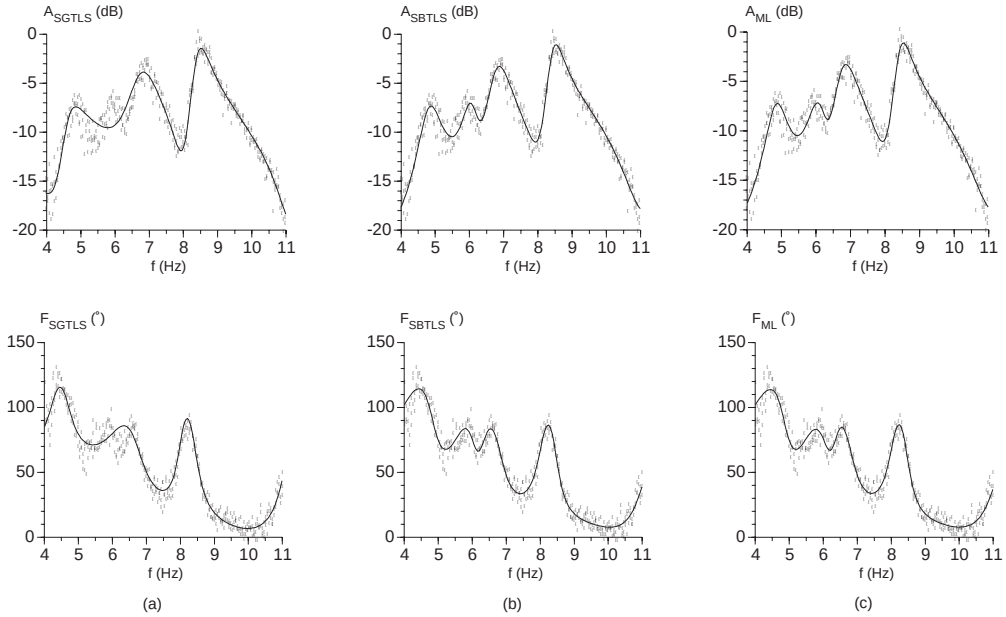


FIG. 3. *Comparison between the measurements (dots) and the estimates (solid lines) of the flight flutter data* (*model, on* = 11, *od* = 10) : (a) *SGTLS,* (b) *SBTLS, and* (c) *ML. From top to bottom: amplitude and phase.*

The big advantage of WGTLS estimators over the optimal ML estimator is that they do not require a nonlinear (iterative) minimization to calculate the solution. In this paper, it has been shown that their disadvantages, poor efficiency, and impracticality since the noise covariance matrix is mostly unknown, can easily be eliminated. Moreover, while the ML estimator using the sample noise covariance matrix needs at least four repeated independent experiments to generate consistent estimates [18], two are sufficient within a TLS framework.

Appendix 1. In this appendix it will be shown that the WGTLS estimation problem (12), after elimination of $\hat{A}$, is equivalent to (13), i.e., both cost functions have the same stationary points. The proof relies on the use of the method of the Lagrangian multipliers. The constrained minimization problem (12) can be reformulated as follows:

$$\begin{aligned} &\arg\min_{\hat{A},X,A} \operatorname{trace}(W[A-\hat{A}][C^TC]^{-1}[A-\hat{A}]^TW^T) + \operatorname{trace}(\Lambda^T\hat{A}X) \\ &\text{subject to } X^TX = I_d, \end{aligned} \tag{63}$$

where $\Lambda \in \mathcal{R}^{n\times d}$ is a Lagrangian multiplier matrix. The use of $\operatorname{trace}(\Lambda^T\hat{A}X)$ is just a convenient way of summing $\Lambda_{ij}[\hat{A}X]_{ij}$ over all i, j. In its minima, the above cost function must be stationary w.r.t. $\hat{A}$, X, and Λ. The derivative w.r.t. $\hat{A}$ yields

$$-2[W^TW][A-\hat{A}][C^TC]^{-1} + \Lambda X^T = 0 \quad \text{or} \quad 2[W^TW][A-\hat{A}] = \Lambda X^TC^TC. \tag{64}$$

Right multiplication of (64) by X, taking into account that $\hat{A}X = 0$, gives

$$\Lambda = 2[W^TW]AX[X^TC^TCX]^{-1}. \tag{65}$$

Elimination of Λ in (64) gives

$$A - \hat{A} = AX[X^T C^T C X]^{-1} X^T C^T C. \tag{66}$$

Replacing $A - \hat{A}$ in (63) by (66), and taking into account the constraint $\hat{A}X = 0$, results in (13). During the proof we have assumed that $W \in \mathcal{R}^{m \times m}$, as well as $C \in \mathcal{R}^{n \times n}$, are nonsingular. It is worthwhile to mention here that even when C is singular the WGTLS solution (13) remains well defined.

Appendix 2. For a common denominator model it is easy to see that $\text{vec}(X^T) = Lx$, where L is a matrix containing ones and zeroes only and where the vector x contains the minimum number of variables to represent all entries of X.

For any parametrization where one or more coefficients are fixed in X, we can still write $\text{vec}(X^T) = L[1, \tilde{x}]^T$, where the vector $\tilde{x}$ stands for the unknowns of X and where a one is included to allow that coefficients of X are known. It boils down to the previous case by introducing the augmented vector $x = [1, \tilde{x}]^T$ and imposing the constraint $x^T x = 1$. The original constraints on X are recovered by appropriate scaling of x after identification: $\tilde{x} = [x_2, x_3, \ldots]^T / x_1$.

REFERENCES

[1] Z. Bai and J. Demmel, *Computing the generalized singular value decomposition*, SIAM J. Sci. Statist. Comput., 14 (1993), pp. 1464–1486.

[2] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, Wiley, London, 1974.

[3] J. W. Brewer, *Kronecker products and matrix calculus in system theory*, IEEE Trans. Circuits and Systems, 25 (1978), pp. 772–781.

[4] E. O. Brigham, *The Fast Fourier Transform*, Prentice–Hall, Englewood Cliffs, NJ, 1974.

[5] D. R. Brillinger, *Time Series: Data Analysis and Theory*, McGraw–Hill, New York, 1981.

[6] P. Eykhoff, *System Indentification, Parameter and State Estimation*, Wiley, New York, 1974.

[7] G. E. Forsythe, *Generation and use of orthogonal polynomial for data-fitting with a digital computer*, J. Industrial Soc. Appl. Math., 5 (1957), pp. 74–88.

[8] 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.

[9] P. Guillaume, R. Pintelon, and J. Schoukens, *A weighted total least squares estimator for multivariable systems with nearly maximum likelihood properties*, in Proc. IEEE Instrumentation and Measurement Technology Conference, IEEE Computer Society Press, Los Alamitos, CA, 1996, pp. 423–428.

[10] P. Guillaume, R. Pintelon, and J. Schoukens, *Accurate estimation of multivariable frequency response functions*, in Proc. 13th Intl. Federation of Automatic Control (IFAC) Triennial World Conference, San Francisco, CA, 1996, pp. 423–428.

[11] P. Guillaume, R. Pintelon, and J. Schoukens, *Parametric identification of multivariable systems in the frequency domain—a survey*, in Proc. ISMA21: International Seminar on Noise and Vibration Engineering, Leuven, Belgium, 1996, pp. 1069–1082.

[12] T. Kailath, *Linear Systems*, Prentice–Hall, New York, 1980.

[13] C. C. Paige, *Computing the generalized singular value decomposition*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 1126–1146.

[14] R. Pintelon, P. Guillaume, Y. Rolain, J. Schoukens, and H. Van hamme, *Parametric identification of transfer functions in the frequency domain, a survey*, IEEE Trans. Automat. Control, 39 (1994), pp. 2245–2260.

[15] Y. Rolain, *Generating robust starting values for frequency domain transfer function estimation*, in Proc. Intl. Federation of Automatic Control (IFAC '96) World Congress, Vol. J, San Francisco, CA, 1996, pp. 173–177.

[16] Y. Rolain, R. Pintelon, K. Q. Xu, and H. Vold, *Best conditioned parametric identification of transfer function models in the frequency domain*, IEEE Trans. Automat. Control, 40 (1995), pp. 1954–1960.

[17] Y. Rolain, G. Vandersteen, D. Schreurs, and S. Van den Bosch, *Parametric modelling of linear time invariant S-parameter devices in the Laplace domain*, in Proc. IEEE Instru-

mentation and Measurement Technology Conference, Brussels, Belgium, IEEE Computer Society Press, Los Alamitos, CA, 1996, pp. 1244–1249.

[18] J. Schoukens, G. Vandersteen, R. Pintelon, and P. Guillaume, *Frequency domain system identification using non-parametric noise models estimated from a small number of data sets*, Automatica IFAC, 33 (1997), pp. 1073–1086.

[19] W. F. Stout, *Almost Sure Convergence*, Academic Press, New York, 1974.

[20] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, Vol. 1, 5th ed., Charles Griffin, London, 1987.

[21] J. Swevers, B. De Moor, and H. Van Brussel, *Stepped sine system identification, error-in-variables and the quotient singular value decomposition*, Mech. Systems Signal Process., 6 (1992), pp. 121–134.

[22] H. Van der Auweraer and J. Leuridan, *Multiple input orthogonal polynomial parameter estimation*, Mech. Systems Signal Process., 1 (1987), pp. 259–272.

[23] G. Vandersteen, H. Van hamme, and R. Pintelon, *General framework for asymptotic properties of generalized weighted nonlinear least squares estimators with deterministic and stochastic weighting*, IEEE Trans. Automat. Control, 41 (1996), pp. 1501–1507.

[24] H. Van hamme and R. Pintelon, *Application of the bootstrapped total least squares (BTLS) estimator in linear system identification*, in Signal Processing VI: Theory and Applications, J. Vandewalle et al., eds., Elsevier Science, Amsterdam, The Netherlands, 1992, pp. 731–734.

[25] S. Van Huffel, H. Park, and J. Ben Rosen, *Formulation and solution of structured total least norm problems for parameter estimation*, IEEE Trans. Signal Proc., 44 (1996), pp. 2464–2474.

[26] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, Philadelphia, PA, 1991.

SIAM J. Matrix Anal. Appl.
Vol. 19, No. 4, pp. 1005–1026, October 1998

011

ON CONSTRUCTION OF A FAMILY OF SMOOTH NONSEPARABLE PREWAVELETS VIA INFINITE PRODUCTS OF TRIANGULARIZABLE MATRICES*

MOHSEN MAESUMI†

Abstract. Infinite products of matrices arise in many areas, such as the study of subdivision and interpolation schemes, Markov chains, and construction of wavelets of compact support. These products are used here to give sufficient conditions for the continuity and differentiability of a class of rectangular compactly supported nonseparable N-dimensional prewavelets or scaling functions. This paper considers the dilation equation $\phi(X) = \sum_K C_K \phi(2X - K)$, where $K \in \{0, \ldots, m\}^N$, $\phi : \mathcal{R}^N \to \mathcal{R}$, and $C_K \in \mathcal{R}$. First, the one-dimensional case is studied, and sufficient conditions on C_K, which guarantee a continuous scaling function $\phi(X)$, are given. These conditions are based on simultaneous triangularizability of two special matrices with entries in terms of C_K. Then, these results are generalized to N dimensions and applied to the particular case where C_K's are obtained by binomial interpolation of their values at the corners of the N-cube, $\{0, m\}^N$. A set of inequalities, based on sums of C_K's on the corners of various faces of the N-cube gives sufficient conditions for the existence of smooth solutions to the dilation equation.

Key words. higher-dimensional scaling functions, infinite matrix products, simultaneous triangularizability.

AMS subject classifications. 26A15, 26A18, 41A05

PII. S0895479894262327

1. Introduction.

Infinite products of matrices occur in a wide variety of fields. They may be used to study subdivision algorithms [2, 13, 14], Markov chains [3], lattice two-scale difference equations [8, 9], and orthonormal bases of compactly supported wavelets [6].

Our interest is in characterizing certain classes of smooth compactly supported N-dimensional prewavelets or scaling functions using infinite products of matrices. These functions are the solutions of the N-dimensional dilation equations

$$\phi(X) = \sum_K C_K \phi(2X - K), \tag{1.1}$$

where $\phi : \mathcal{R}^N \to \mathcal{R}$, $C_K \in \mathcal{R}$, $K \in \{0, \ldots, m\}^N$, and $X \in \mathcal{R}^N$.

If the values of ϕ at integer points are known, then one can use (1.1) to get the values of ϕ at half integers and, by iterating the process, at all dyadic points (i.e., at $\mathcal{Z}^N/2^\ell$ for all nonnegative ℓ). Since any X can be approximated by a sequence of dyadic points, e.g., through its binary expansion, the continuity of ϕ will then provide the value of $\phi(X)$.

An efficient way to describe this process is to convert (1.1) to a matrix equation. Since the support of the solution is finite, the matrix will be finite too. Then, the iteration will take the form of multiplication by certain fixed matrices with entries in terms of C_K. There is a one-to-one correspondence between the digits of the binary expansion of X and the matrices that appear in the product. As more digits of the

*Received by the editors January 28, 1994; accepted for publication (in revised form) by P. Van Dooren July 8, 1997; published electronically July 17, 1998. This research was supported in part by Texas Advanced Research Program grant 003581-005.
http://www.siam.org/journals/simax/19-4/26232.html
†Lamar University, P. O. Box 10047, Beaumont, TX, 77710 (maesumi@math.lamar.edu).

binary expansion are taken into account, the length of the matrix product increases. Hence, we will take up the question of infinite products of matrices.

In section 2 of this paper, we study the one-dimensional case. Our main result in that section, Theorem 2.5, classifies a one-parameter class of $\mathcal{C}^\ell$ scaling functions for $\ell < m-1$. In section 3, we generalize our results to the N-dimensional case. In Theorem 3.3, we classify a (2^N-1)-parameter family of $\mathcal{C}^\ell$ scaling functions in N dimensions.

2. One-dimensional scaling functions. In one dimension, the dilation equation may be written as

$$\phi(x) = c_0\phi(2x) + c_1\phi(2x-1) + \cdots + c_m\phi(2x-m), \tag{2.1}$$

where $\phi : \mathcal{R} \to \mathcal{R}$ and c_i, $i = 0, \ldots, m$, are given real coefficients. The regularity properties of the solutions of dilation equations have been extensively studied. In particular, nontrivial $\mathcal{L}^1$ solutions having compact support are characterized in [8] and shown to have their support in $[0, m]$. Moreover, it is shown that if ϕ is r times continuously differentiable, then $r < m-1$. The Hölder exponent and fractal structure of ϕ are determined in [4, 5, 9]. Continuous solutions are characterized in terms of the general and joint spectral radii of a family of matrices in [10] (see also [1, 11, 12, 15, 16]).

The point of view in the next section of this paper is to identify certain smooth one-dimensional scaling functions which lead to the specification of certain smooth solutions in higher dimensions. Some higher-dimensional scaling functions can be formed by tensor products of lower-dimensional ones. Our solution is, different however, and cannot be reduced to a tensor product.

Our construction depends on results concerning infinite products of matrices. Given a pair of matrices, T_0 and T_1, any infinite product (e.g., $P = T_0T_1T_1T_0T_1\cdots$) is associated with a binary number (e.g., $x = .01101\ldots$). We give sufficient conditions for (a) the convergence of such products; (b) the existence of a well-defined map that, given any $x \in [0,1]$, generates a product; and (c) the continuous dependence of the product on x. The sufficient conditions require that (I) the two matrices are simultaneously triangularizable by a similarity transformation, (II) the first diagonal elements of triangular matrices are 1 and the remaining elements are less than one in absolute value, and (III) the products of each matrix with the eigenvector of the other matrix (associated with eigenvalue 1) are linearly dependent. While considerably weaker conditions that guarantee the same results are known (see [10]), our requirement of simultaneous triangularizability can be easily adapted to identify certain continuous prewavelets in higher dimensions. In particular, we will characterize a (2^N-1)-parameter family of continuous $(m+1)^N$-coefficient scaling functions in N dimensions. Similar results are obtained for higher-order regularity.

2.1. Notation. Define the vector Φ and matrices T_0 and T_1 by

$$\Phi(x) = [\phi(x), \phi(x+1), \ldots, \phi(x+m-1)]^t \quad \text{for} \quad 0 \le x \le 1, \tag{2.2a}$$

$$c_k = 0 \quad \text{for} \quad k < 0 \quad \text{or} \quad k > m, \tag{2.2b}$$

$$(T_d)_{ij} = c_{2i-j+d-1} \quad \text{for} \quad 1 \le i, j \le m \quad \text{and} \quad d = 0 \text{ or } 1, \tag{2.2c}$$

$$T_0 = \begin{pmatrix} c_0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ c_2 & c_1 & c_0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ c_4 & c_3 & c_2 & c_1 & c_0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & c_m & c_{m-1} & c_{m-2} & c_{m-3} \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & c_m & c_{m-1} \end{pmatrix},$$

$$T_1 = \begin{pmatrix} c_1 & c_0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ c_3 & c_2 & c_1 & c_0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ c_5 & c_4 & c_3 & c_2 & c_1 & c_0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & c_m & c_{m-1} & c_{m-2} \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & c_m \end{pmatrix}.$$

(In what follows, the range of i and j is $\{1, 2, \ldots, m\}$ unless further restricted.) Notice that the definition of Φ is based on dividing the interval $[0, m]$ into m *cells*, $[i-1, i]$, $1 \le i \le m$. We define a pair of vectors or a function $f : \{0, 1\} \to \mathcal{R}^m$ to be *shift continuous* if $f(0)_i = f(1)_{i-1}$ for $1 < i \le m$. Obviously, $\phi(x)$ is continuous on $[0, m]$ iff Φ is continuous on $[0, 1]$ and is shift continuous.

We search for the unique normalized continuous solution ϕ with support in $[0, m]$. Sufficient conditions for the existence of such a solution and particular examples are given in the theorems of this section. This solution satisfies $\phi(x) = 0$ for $x \le 0$ or $x \ge m$, $\Phi(0)_1 = \Phi(1)_m = 0$, and

$$\Phi(x) = T_{x_1}\Phi(2x - x_1), \tag{2.3}$$

where x_1 is the first digit in the binary expansion of x. In particular if we apply (2.3) to $x = 0 = 0.00\ldots$, $x = 1 = 0.11\ldots$, and $x = 1/2 = 0.100\ldots = 0.011\ldots$, respectively, then we get

$$T_0\Phi(0) = \Phi(0), \qquad T_1\Phi(1) = \Phi(1), \tag{2.4a}$$

$$T_0\Phi(1) = T_1\Phi(0) = \Phi(1/2). \tag{2.4b}$$

Once $\Phi(0)$ or $\Phi(1)$ is known, one can calculate Φ at dyadics by repeated applications of (2.3).

Now, suppose $x \in [0, 1]$ and indicate its binary expansion by $x = 0.x_1x_2\cdots x_qx_{q+1}\cdots$. Denote by $\bar{x}_q$ the residual after the qth digit, $\bar{x}_q = 0.x_{q+1}x_{q+2}\cdots$. Then, by repeated application of (2.3), we get

$$\Phi(x) = \prod_{\ell=1}^{q} T_{x_\ell}\Phi(\bar{x}_q). \tag{2.5}$$

We define $P_q(T_0, T_1, x) = \prod_{\ell=1}^{q} T_{x_\ell}$ and $P(T_0, T_1, x) = \lim_{q\to\infty} P_q(T_0, T_1, x)$ whenever the limit exists.

Dyadic numbers have two binary expansions, e.g., $x = 1/2 = .100\ldots = .011\ldots$. Therefore, in the definition of $P_q(T_0, T_1, x)$ a particular expansion of x should be specified a priori. The consistency of (2.3) at dyadics, i.e., (2.4b), remedies this

nonuniqueness for the infinite products, and the value of $P(T_0, T_1, x)$ is then determined independently of the choice of expansion for x. Further details are provided in Lemmas 2.4 and 2.5 below.

The matrices T_0 and T_1 have a very special structure. For example, the submatrices obtained by deleting the first row and column of T_0 are the same as the one obtained by deleting the last row and column of T_1. Moreover, the columns of T_0 and T_1 contain all the c_k's with an even index or all the c_k's with an odd index. However, these special properties are not used before Lemma 2.7. For this reason, and to simplify the notation, we use matrices A and B in place of T_0 and T_1, respectively.

2.2. Conditions for convergence of P_q. Some of the elementary necessary conditions for existence of $P(A, B, x)$ are expressed in the following lemma.

LEMMA 2.1. *Let Q be a finite product of A's and B's and λ be an eigenvalue of Q. Then, $P(A, B, x)$ exists for all $0 \le x \le 1$ only if $|\lambda| < 1$ or $\lambda = 1$ and nondefective.*

Proof. These conditions follow immediately from the Jordan normal form of the matrix under consideration. Here, we give a brief indication. If an eigenvalue $|\lambda| > 1$, then Q^ν is exponentially unbounded as $\nu \to \infty$, and the corresponding product does not exist. If $\lambda = 1$ is defective (i.e., its geometric multiplicity is less than its algebraic multiplicity), then Q^ν is polynomially unbounded. If $|\lambda| = 1$ but $\lambda \neq 1$, then Q^ν does not have a limit. □

We note that if all the eigenvalues of Q are less than 1 in absolute value, then $P(A, B, x)$ will be zero on a dense set of values of x. (To see this, consider the set of numbers whose binary expansions end in an infinite repetition of the digit pattern associated with Q. These numbers form a dense set and P is zero on this set.) One of the simplest cases for controlling the eigenvalues of products of matrices is when the matrices are triangular. This prompts the following definition.

DEFINITION 2.2. *A finite family of matrices $\{\mathcal{A}\}$ is said to be jointly tied (to 1) if the matrices are simultaneously lower triangularizable by a similarity transformation, their leading eigenvalue is 1, and the remaining eigenvalues are less than 1 in absolute value. Hence, there is an invertible matrix S such that for each $A \in \mathcal{A}$ we have*

$$S^{-1}AS = \tilde{A}, \quad \tilde{A}_{ij} = 0 \quad \text{for} \quad j > i,$$

$$\tilde{A}_{11} = 1, \quad \text{and} \quad \max_{i>1} |\tilde{A}_{ii}| < 1.$$

The following definition will be used to study the relationship between the two products associated with the two expansions of the dyadics.

DEFINITION 2.3. *Two matrices A and B are called consistent (with respect to a simple joint eigenvalue λ) if there are V_A and V_B, eigenvectors of A and B associated with λ, such that $AV_B = BV_A$.*

THEOREM 2.4. *Let A and B be jointly tied. Then, for a given binary expansion of x, $P(A, B, x)$ exists. P is continuous at x if x is nondyadic. If A and B are jointly tied and consistent with respect to the joint eigenvalue 1, then $P(A, B, x)$ is well defined and continuous for all x.*

We establish this theorem by proving Lemmas 2.2 through 2.5.

LEMMA 2.5. *Let U be an $m \times m$ lower triangular matrix with $U_{11} = 1$, $|U_{ii}| < 1$ for $i > 1$, and $U_{ii} \neq U_{jj}$ for $i \neq j$. Then, $\lim_{\nu\to\infty} U^\nu = U^\infty$ exists, and its nonzero entries are only on the first column. Moreover, ${U^\nu}_{ij} \to 0$ exponentially for $j > 1$.*

Proof. The eigenvalues of U, $\{1, U_{22}, \ldots, U_{mm}\}$, are distinct; hence, U is diagonalizable by a similarity transformation. We write $U = S\Delta S^{-1}$, where Δ is diagonal, S and S^{-1} are lower triangular, and $\Delta_{11} = S_{11} = {S^{-1}}_{11} = 1$. Now, $U^\infty = S\Delta^\infty S^{-1}$, ${\Delta^\infty}_{11} = 1$, ${\Delta^\infty}_{ij} = 0$ for $(i,j) \neq (1,1)$. As S and S^{-1} are lower triangular, we get ${U^\infty}_{i1} = S_{i1}$, and the remaining elements of U^∞ are zero. Let $\epsilon = \max_{i>1} |U_{ii}|$. Then from $U^\nu = S\Delta^\nu S^{-1}$ we get ${U^\nu}_{ij} = O(\epsilon^\nu)$ for $j > 1$. □

Remark 2.1. The convergence ${U^\nu}_{ij} \to U^\infty$ for $j > 1$ will occur even if U_{ii} for $i > 1$ are not distinct. This is evident from the Jordan normal form of U. The convergence rate, however, could be slower. If the largest Jordan block associated with ϵ is of size q, then the elements of ${U^\nu}_{i,j}$ for $j > 1$ are at most of the order of $\binom{\nu}{q-1}\epsilon^{\nu-q+1}$.

Remark 2.2. If U is triangular and for a fixed i and any $j > i$ we have $|U_{ii}| > |U_{jj}|$, then, as $\nu \to \infty$, the ith column of $(U/U_{ii})^\nu$ converges to a finite vector and all subsequent columns tend to zero.

LEMMA 2.6. *Let A and B be jointly tied. Then, for a given binary expansion of x, $\tilde{P} = P(\tilde{A}, \tilde{B}, x)$ exists and the only nonzero entries of $\tilde{P}$ are in its first column. In particular, $\tilde{P}_{11} = 1$. $\tilde{P}$, as a function of x, is uniformly bounded.*

Proof. Let $M = \max_{i,j}\{|\tilde{A}_{ij}|, |\tilde{B}_{ij}|\}$ and $\epsilon = \max_{i>1}\{|\tilde{A}_{ii}|, |\tilde{B}_{ii}|\}$. Choose δ and $m-1$ distinct ϵ_i's such that $\epsilon < \epsilon_i < \delta < 1$ for $i > 1$. Define a lower triangular matrix U with $U_{11} = 1$, $U_{ii} = \epsilon_i$ for $i > 1$, $U_{ij} = M$ for $i > j$, and $U_{ij} = 0$ for $j > i$. The absolute values of entries of $\tilde{A}$ and $\tilde{B}$ are dominated by those of U; hence, $|P_q(\tilde{A}, \tilde{B}, x)_{ij}| \leq {U^q}_{ij}$. Now, by Lemma 2.2, U^q converges to a matrix whose nonzero elements are on its first column only. Therefore, $\lim_{q\to\infty} P_q(\tilde{A}, \tilde{B}, x)_{ij} = 0$ for $j > 1$. Moreover, as $q \to \infty$, $P_q(\tilde{A}, \tilde{B}, x)_{i1}$ appears as a series with exponentially decaying terms; hence, it converges. Specifically, denote $P_q(\tilde{A}, \tilde{B}, x)$ by $\tilde{P}_q$, and write $\tilde{P}_q - \tilde{P}_1 = \sum_{\ell=1}^{q-1} \tilde{P}_{\ell+1} - \tilde{P}_\ell = \sum_{\ell=1}^{q-1} \tilde{P}_\ell(\tilde{D} - I)$, where I is identity, $\tilde{D} = \tilde{A}$ if $x_{\ell+1} = 0$, and $\tilde{D} = \tilde{B}$ if $x_{\ell+1} = 1$. Now, $|(\tilde{D} - I)_{ij}| < M + 1$ and the exponential decay of ${U^q}_{ij}$ for $j > 1$ implies $(\tilde{P}_\ell)_{ij} = o(\delta^\ell)$ for $j > 1$. Using $(\tilde{D} - I)_{11} = 0$ we get $(\tilde{P}_\ell(\tilde{D} - I))_{ij} = o(\delta^\ell)$ for all i and j. Therefore, $\tilde{P}_q$ converges as $q \to \infty$. We have $(\tilde{P}_q)_{11} = 1$ for all q, and hence $\tilde{P}_{11} = 1$. Note that $\tilde{P}(\tilde{A}, \tilde{B}, x)$ is uniformly bounded by U^∞ for all x. □

Define Z to be the first column of $\tilde{P}$, $Z_i = P(\tilde{A}, \tilde{B}, x)_{i1}$. Note that $\tilde{P}Z = Z$ and $Z_1 = 1$. Let $W = SZ$. From $P = S\tilde{P}S^{-1}$ we get $P_{ij} = W_i {S^{-1}}_{1j}$ and $PW = W$; i.e., W is the eigenvector of P associated with eigenvalue 1. (All other eigenvalues are zero and the null space is generated by columns 2 through m of S.)

LEMMA 2.7. *Let A and B be jointly tied and consistent; then $AB^\infty = BA^\infty$.*

Proof. Since A and B are jointly tied, then, by Lemma 2.3, A^∞ and B^∞ exist. Moreover, similarity transformation preserves consistency, and $\tilde{A}$ and $\tilde{B}$ are also consistent. We have $AB^\infty - BA^\infty = S(\tilde{A}\tilde{B}^\infty - \tilde{B}\tilde{A}^\infty)S^{-1}$. Now, $\tilde{B}^\infty$ has only zeros on columns 2 through m, and the first column is just the eigenvector of $\tilde{B}$ whose first entry is 1. The same applies to $\tilde{A}$. We have $(\tilde{A}\tilde{B}^\infty - \tilde{B}\tilde{A}^\infty)_{i,j} = 0$ because, for $j > 1$, ${B^\infty}_{ij} = {A^\infty}_{ij} = 0$, and, for $j = 1$, the cancellations occur due to consistency of $\tilde{A}$ and $\tilde{B}$. □

LEMMA 2.8. *Let A and B be jointly tied and consistent; then $P(A, B, x)$ is a continuous function of x.*

Proof. We prove this first for the case when x is not dyadic and then for the case when x is dyadic. Only in the latter case do we use the consistency of A and B. The similarity transformation preserves continuity. Hence, it is sufficient to prove that $P(\tilde{A}, \tilde{B}, x)$ is a continuous function of x.

Case 1. Assume that x is not dyadic. Then, the binary expansion of x does not have a tail of zeros or a tail of ones. Hence, $y \to x$ implies that an increasing number of digits of y agree with those of x.

Suppose that y agrees with x on the first q digits; then

$$P(\tilde{A},\tilde{B},x) - P(\tilde{A},\tilde{B},y) = P_q(\tilde{A},\tilde{B},x)[P(\tilde{A},\tilde{B},\bar{x}_q) - P(\tilde{A},\tilde{B},\bar{y}_q)]. \tag{2.6a}$$

Now, for sufficiently large q, $P_q(\tilde{A},\tilde{B},x)$ has near zero entries in positions (i,j) for $j > 1$. Moreover, $[P(\tilde{A},\tilde{B},\bar{x}_q) - P(\tilde{A},\tilde{B},\bar{y}_q)]$ has a zero entry in the $(1,1)$ position and the remaining entries are uniformly bounded. As a result, the right-hand side of (2.6a) approaches zero as $q \to \infty$. Therefore, we have $\lim_{y\to x} P(\tilde{A},\tilde{B},y) = P(\tilde{A},\tilde{B},x)$ for x nondyadic.

Case 2. Assume that x is dyadic. If y approaches x while agreeing with an increasing number of digits of x, then Case 1 applies. Otherwise let $x = 0.x_1x_2\cdots x_q1000\cdots$ and $y = 0.x_1x_2\cdots x_q01\cdots 1y_{q+\nu+2}y_{q+\nu+3}\cdots$, where the ν digits y_{q+2} through $y_{q+\nu+1}$ are equal to 1. Note that $y \to x$ as $\nu \to \infty$, but only the first q digits of y and x agree.

Now, we write

$$P(\tilde{A},\tilde{B},x) - P(\tilde{A},\tilde{B},y) = P_q(\tilde{A},\tilde{B},x)[\tilde{B}\tilde{A}^\infty - \tilde{A}\tilde{B}^\nu P'], \tag{2.6b}$$

where $P' = P(\tilde{A},\tilde{B},\bar{y}_{q+\nu+1})$. Notice that $\lim_{\nu\to\infty}\tilde{B}^\nu P' = \tilde{B}^\infty$ since all columns of $\tilde{B}^\nu$, except the first one, approach zero while $P'_{11} = 1$ and P' stays uniformly bounded. Lemma 2.4 gives $\lim_{\nu\to\infty}\tilde{B}\tilde{A}^\infty - \tilde{A}\tilde{B}^\nu P' = \tilde{B}\tilde{A}^\infty - \tilde{A}\tilde{B}^\infty = 0$. Therefore, $\lim_{y\to x} P(\tilde{A},\tilde{B},y) = P(\tilde{A},\tilde{B},x)$ for x dyadic. □

This concludes the proof of Theorem 2.1.

A function f is said to have Hölder exponent (at least) α for $0 \le \alpha \le 1$ if there is $C \ge 0$ such that $|f(x) - f(y)| \le C|x-y|^\alpha$. Based on this definition, we can obtain additional regularity information about P by combining (2.6a) and (2.6b).

LEMMA 2.9. *Suppose* $1 > 2^{-r} = \delta > \max_{i>1}\{|\tilde{A}_{ii}|, |\tilde{B}_{ii}|\}$. *Then the Hölder exponent of* P *is at least* $r = -\log_2(\delta)$.

Proof. Consider $1 \ge y > x \ge 0$; then the binary expansions of x and y will be $x = 0.x_1\cdots x_q011\cdots 1x_{q+n+2}x_{q+n+3}\cdots$ and $y = 0.x_1\cdots x_q100\cdots 0y_{q+n+2}y_{q+n+3}\cdots$, where the first q digits are identical and the digits in positions $q+2$ through $q+n+1$ are ones for x and zeros for y. Moreover $x_{q+n+2} = y_{q+n+2}$ or $x_{q+n+2} = 0$ and $y_{q+n+2} = 1$. In the former case $y - x \ge 2^{-(q+n+2)}$, and in the latter case $y - x \ge 2^{-(q+n+1)}$. In either case $|y-x|^r \ge \delta^{n+q+2}$. We have $\tilde{P}(y) - \tilde{P}(x) = \tilde{P}_q(x)[\tilde{B}\tilde{A}^n\tilde{P}(\bar{y}_{q+n+1}) - \tilde{A}\tilde{B}^n\tilde{P}(\bar{x}_{q+n+1})]$. The absolute values of the (i,j) elements of $\tilde{P}(q)$ for $j > 1$ are bounded by $C_1\delta^q$ for some $C_1 > 0$. The $(1,1)$ entry of the bracket is zero, and all others are bounded by $C_2\delta^n$. Hence $|\tilde{P}(y) - \tilde{P}(x)| \le C_1C_2\delta^{n+q} \le C|y-x|^r$, where $C = C_1C_2\delta^{-2}$. Hence the Hölder exponent of $\tilde{P}(x)$, and therefore that of $P(x)$, is at least $r = -\log_2(\delta)$. □

We have identified sufficient conditions for $P(A,B,x)$ to be well defined and continuous. Now we concentrate on the special matrices given by (2.2c). The following two lemmas make full use of the particular structure of T_0 and T_1. They will be used to specialize the result of Theorem 2.1 to the solution of the dilation equation (2.1).

LEMMA 2.10. *Let* G *be the matrix obtained by removing the first row and column of* T_0 *or the last row and column of* T_1. *Thus* $G_{ij} = c_{2i-j}$, $1 \le i,j \le m-1$. *Suppose that* G *has a right eigenvector* $V = (g_1, g_2, \ldots, g_{m-1})$ *associated with eigenvalue* α; *then* T_0 *has a right eigenvector* $V_0 = (0, g_1, g_2, \ldots, g_{m-1})$ *and* T_1 *has a right*

eigenvector $V_1 = (g_1, g_2, \ldots, g_{m-1}, 0)$*, both with the same eigenvalue* α *and satisfying* $T_0V_1 = T_1V_0$*. If* α *is simple, then* T_0 *and* T_1 *are consistent with respect to* α*. If* $\sum_k c_{2k} = \sum_k c_{2k+1} = \beta$*, then* G*,* T_0*, and* T_1 *each have a left eigenvector of the form* $(1, 1, \ldots, 1)$ *with the same eigenvalue* β*. Suppose that* β *is simple; then* T_0 *and* T_1 *are consistent with respect to the corresponding right eigenvectors and* β*.*

Proof. This is immediate from the special structure of T_0 and T_1. □

DEFINITION 2.11. *The coefficients* c_k *are said to satisfy the unit column sum rule if*

$$\sum_k c_{2k} = \sum_k c_{2k+1} = 1. \tag{2.7}$$

If c_k satisfy the unit column sum rule and 1 is a simple eigenvalue, then T_0 and T_1 will be consistent with respect to 1. If, in addition, T_0 and T_1 are jointly tied, then our construction yields a continuous solution of (2.3) and the corresponding continuous solution of (2.1). (Notice that $\Phi(0)$ and $\Phi(1)$ are eigenvectors of T_0 and T_1 corresponding to eigenvalue 1. According to Lemma 2.7, they are shift continuous, i.e., $\Phi(1)_{i-1} = \Phi(0)_i$ for $i > 1$.) Now, we proceed to show that ϕ is properly normalized.

LEMMA 2.12. *Let* ϕ *be a continuous solution of* (2.1)*, and assume* $\Gamma = \int \phi(x)dx \neq 0$*. Then*

$$\sum_k c_k = 2. \tag{2.8}$$

If $\Gamma = 1$*, then*

$$\sum_k c_{2k} \sum_n \phi(2n+1) + \sum_k c_{2k+1} \sum_n \phi(2n) = 1. \tag{2.9}$$

Moreover, if $\Gamma = 1$ *and* c_k *satisfy the unit column sum rule* (2.7)*, then for any* x

$$\sum_k \phi(k+x) = 1. \tag{2.10}$$

Proof. The first sum rule (2.8) for c_k's is obtained by integrating (2.1). (We use the compactness of the support of ϕ and c_k to simplify our formulas. Unless otherwise indicated, the integrals are over the entire reals and the summations are over the entire integers.) To establish (2.9) we form a Riemann sum for the integral and simplify the sum using (2.1).

Consider the dyadics points at a fixed level ℓ, i.e., the ones of form $(2n+1)/2^\ell$. We use these points to form a Riemann sum S_ℓ to approximate $\int \phi$. We have $S_\ell = 2^{1-\ell} \sum_n \phi((2n+1)/2^\ell)$. Now, we apply the recursion relation (2.1) to write $\phi((2n+1)/2^\ell)$ in terms of the dyadics at level $\ell - 1$. Assume $\ell > 1$; then we have

$$\begin{aligned}\sum_n \phi\left(\frac{2n+1}{2^\ell}\right) &= \sum_n \sum_k c_k \phi\left(\frac{2n+1}{2^{\ell-1}} - k\right) \\ &= \sum_k c_k \sum_n \phi\left(\frac{2n+1-k2^{\ell-1}}{2^{\ell-1}}\right) = \sum_k c_k \sum_n \phi\left(\frac{2n+1}{2^{\ell-1}}\right).\end{aligned}$$

Therefore, $S_\ell = 1/2 \sum_k c_k S_{\ell-1}$. Hence, if $\ell > 1$ and $\sum c_k = 2$, then $S_\ell = S_{\ell-1}$. However, if $\ell = 1$, we get

$$\begin{aligned} S_1 &= \sum_n \phi\left(\frac{2n+1}{2}\right) = \sum_n \sum_k c_k \phi(2n+1-k) \\ &= \sum_k c_{2k} \sum_n \phi(2n+1-2k) + \sum_k c_{2k+1} \sum_n \phi(2n+1-(2k+1)) \\ &= \sum_k c_{2k} \sum_n \phi(2n+1) + \sum_k c_{2k+1} \sum_n \phi(2n). \end{aligned}$$

Hence, $S_\ell = S_1 = \sum_k c_{2k} \sum_n \phi(2n+1) + \sum_k c_{2k+1} \sum_n \phi(2n)$. Now, as $\ell \to \infty$, we have $S_\ell \to \int \phi = 1$, which proves (2.9). Moreover, if c_k satisfy the unit column sum rule, then we get $\sum_\ell \phi(\ell) = 1$. (One uses this result to normalize the eigenvectors of T_0 and T_1 corresponding to eigenvalue 1 in (2.4a). That is, $\sum_j \Phi(0)_j = \sum_j \Phi(1)_j = 1$.)

Finally, we prove (2.10) and show that the integral of ϕ equals the sum of ϕ at any translate of the integers. Consider a vector $V = (v_1, v_2, \ldots, v_m)^t$. From (2.7), one can easily see $\sum_i (T_0 V)_i = \sum_i (T_1 V)_i = \sum_i V_i$. Hence, if we start with $V = \Phi(0)$ or $V = \Phi(1)$ and multiply on the left with T_0's or T_1's, then at any stage the resulting values of $\phi(x)$ at dyadics satisfy (2.10), and in the limit the same equation is satisfied at all points by continuity of ϕ. □

THEOREM 2.13. *If T_0 and T_1 are jointly tied and their entries c_k satisfy the unit column sum rule, then $P(T_0, T_1, x)$ is well defined and continuous, the columns of P are identical, and a solution of* (2.1) *is given by $\phi(x+i-1) = P(T_0, T_1, x)_{i,j}$ for any j. Moreover, this ϕ is properly normalized, i.e., $\int \phi dx = 1$.*

Proof. Since c_k's satisfy the unit column sum rule and 1 is a simple eigenvalue, then, by Lemma 2.7, T_0 and T_1 are consistent. The matrices are assumed to be jointly tied; therefore, by Lemma 2.5, P is well defined and continuous. Since c_k's satisfy the unit column sum rule, then, by the argument in the proof of (2.10), the sum of elements of any column of any product of T_0's and T_1's, e.g., $P(T_0, T_1, x)$, is 1. Now, by the comments following Lemma 2.3, we have $P_{ij} = W_i {S^{-1}}_{1j}$ and $1 = \sum_i P_{ij} = {S^{-1}}_{1j} \sum W_i$ for any j. Hence, the elements of the first row of S^{-1} are equal, and we may assume ${S^{-1}}_{1j} = 1$. Then, the columns of P and W are equal, and each represents ϕ through $\phi(x+i-1) = P(T_0, T_1, x)_{i,j}$ for any j. Using Lemma 2.8 we get $\int \phi = \sum_i \phi(x+i-1) = \sum_i P_{ij} = 1$. Hence, ϕ is properly normalized. □

2.3. Infinite products of a finite family of matrices. Theorem 2.1 can be generalized to include the products of a finite family of matrices. Let $R > 1$ be an integer and r be a digit in base R, i.e., $0 \le r \le R-1$. Consider R matrices A_0, A_1, ..., A_{R-1}. Represent $x \in [0,1]$ by its expansion in base R, $x = 0.x_1 x_2 \cdots$ (now $0 \le x_q \le R-1$). Define $P_q(x) = \prod_{\ell=1}^{q} A_{x_\ell}$ and $P(x) = \lim_{q\to\infty} P_q(x)$ whenever the limit exists. The family is called consistent if there are simple eigenvectors V_0 and V_{R-1} such that $A_0 V_0 = V_0$, $A_{R-1} V_{R-1} = V_{R-1}$, and $A_r V_{R-1} = A_{r+1} V_0$ for $0 \le r \le R-2$. Now, we have the following theorem.

THEOREM 2.14. *Let the family $\{A_r\}$ be jointly tied and consistent. Then, $P(x)$ exists and is continuous.*

Proof. This is similar to Theorem 2.1. □

2.4. Analysis of three-term dilation equations. In this section we give an example based on the case $m = 2$. We can achieve triangularization if $c_1 = c_0 + c_2$,

in which case for any $a \neq b$ we have

$$T_0 = \begin{pmatrix} c_0 & 0 \\ c_2 & c_1 \end{pmatrix} = \frac{1}{a-b}\begin{pmatrix} a & -1 \\ -b & 1 \end{pmatrix}\begin{pmatrix} c_1 & 0 \\ ac_2 & c_0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ b & a \end{pmatrix},$$

$$T_1 = \begin{pmatrix} c_1 & c_0 \\ 0 & c_2 \end{pmatrix} = \frac{1}{a-b}\begin{pmatrix} a & -1 \\ -b & 1 \end{pmatrix}\begin{pmatrix} c_1 & 0 \\ bc_0 & c_2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ b & a \end{pmatrix}.$$

We note that the triangularization is not unique. We set $a = 0$ and $b = 1$ to get

$$T_0 = \begin{pmatrix} c_0 & 0 \\ c_2 & c_1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} c_1 & 0 \\ 0 & c_0 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix},$$

$$T_1 = \begin{pmatrix} c_1 & c_0 \\ 0 & c_2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} c_1 & 0 \\ c_0 & c_2 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}.$$

Now, $\tilde{T}_0$ and $\tilde{T}_1$ are the middle matrices on the right-hand side of previous equations. Given an x it is easy to form $\tilde{P} = P(\tilde{T}_0, \tilde{T}_1, x)$. For $c_1 = 1$, $0 < c_0 < 1$, and $0 < c_2 < 1$, we have $\tilde{P}_{11} = 1$, $\tilde{P}_{12} = \tilde{P}_{22} = 0$. Define $\sigma_q = \sigma_q(x) = \sum_{n=1}^{q} x_n$ and $\sigma_0(x) = 0$. Then, $(\tilde{P}_{q+1})_{21} = (\tilde{P}_q)_{21} + x_{q+1}c_0(\tilde{P}_q)_{22}$ and $(\tilde{P}_q)_{22} = {c_0}^{q-\sigma_q}{c_2}^{\sigma_q}$. Therefore,

$$\tilde{P}_{2,1} = z(x) = \sum_{q=0}^{\infty} x_{q+1}{c_0}^{q+1-\sigma_q}{c_2}^{\sigma_q}.$$

Now, using our previous notation (following Lemma 2.3) we have $Z(x) = (1, z(x))^t$ and

$$P(T_0, T_1, x) = \begin{pmatrix} z(x) & z(x) \\ 1 - z(x) & 1 - z(x) \end{pmatrix}, \quad \phi(x) = \begin{cases} z(x) & \text{for } 0 \le x \le 1, \\ 1 - z(x-1) & \text{for } 1 \le x \le 2. \end{cases}$$

It is easy to verify that $\phi(x)$ is increasing on $[0, 1]$ and decreasing on $[1, 2]$.

2.5. Simultaneous triangularization. The main step in our analysis of the products of two matrices is to reduce them to a triangular form. Given A and B, we search for $\tilde{A}$, $\tilde{B}$, and S^{-1} such that $S^{-1}A = \tilde{A}S^{-1}$ and $S^{-1}B = \tilde{B}S^{-1}$. This constitutes $2m^2$ nonlinear algebraic equations. In the case of wavelets, one always enforces (2.7). Then, ${S^{-1}}_{1j} = \tilde{A}_{11} = \tilde{B}_{11} = 1$. This reduces the number of equations to $2m^2 - 2m$ and the number of unknowns to $2m^2 - 2$. (The eigenvalues of A and B are the elements on the diagonal of $\tilde{A}$ and $\tilde{B}$, but their positions are not known.) Therefore, there are $2m - 2$ degrees of freedom in the triangularization (e.g., a and b in section 2.4). Despite the presence of degrees of freedom, simultaneous triangularization is rarely possible. (It is known that a family of matrices is simultaneously triangularizable iff the eigenvalues of any product of matrices are equal, in some order, to the products of eigenvalues of the same matrices.)

The triangularizer matrices which are useful for the construction of N-dimensional scaling functions are the ones which work for a class of matrices and have constant entries. For example, if $m = 3$, then T_0 and T_1 can be triangularized when $c_0 + c_3 = 1$ or $1/2$. But if the sum is 1, then the triangularizer depends on c_0 and will not be suitable for higher-dimensional constructions considered in this paper. On the other hand when the sum is $1/2$, then the triangularizer is constant. In the next section we focus on the latter case.

2.6. Analysis of $(m+1)$-term dilation equations. A class of matrices T_0 and T_1 for which constant triangularizers have been obtained are exactly those which satisfy certain sum rules used to enforce high regularity [9]. Here we require a particular subset of such rules, i.e.,

$$\sum_k c_k k^q (-1)^k = 0 \quad \text{for} \quad q = 0, \ldots, m-2, \tag{2.11}$$

where 0^0 is taken to be 1. One can solve (2.11) for $c_1, \ldots, c_{m-1}$ in terms of the "corner" values c_0 and c_m. The resulting coefficients c_k and the associated matrices T_0 and T_1 satisfy a host of binomial-type identities. The outlines for the proof of some of these identities are collected in Appendix A as notes. In what follows, we adopt the usual conventions that $\binom{a}{b} = 0$ if $b > a$ or $b < 0$, and $1/c! = 0$ if $c < 0$.

The coefficient matrix of (2.11) is a Vandermonde-type matrix with a nonzero determinant. Therefore, it is nonsingular. The unique solution is given by "binomial interpolation" between the endpoint values (see Note A.1),

$$c_k = c_0 \binom{m-1}{k} + c_m \binom{m-1}{k-1}. \tag{2.12}$$

For this particular choice of c_k's, an $m \times m$ triangularizer matrix S and its inverse S^{-1} are given by (see Note A.2.)

$$S_{ij} = \binom{j-1}{m-i} \frac{(-1)^{i+j-m-1}}{(j-1)!}, \qquad S^{-1}_{ij} = \binom{m-j}{i-1} (i-1)!. \tag{2.13}$$

(Notice that the entries of S and S^{-1} are zero below the second diagonal, i.e., if $i+j > m+1$.) Upon triangularization, the diagonal elements of $\tilde{T}_0$ and $\tilde{T}_1$, respectively, are (see Note A.2)

$$\begin{array}{l} 2^{m-2}(c_0+c_m), 2^{m-3}(c_0+c_m), \ldots, (c_0+c_m), c_0, \\ 2^{m-2}(c_0+c_m), 2^{m-3}(c_0+c_m), \ldots, (c_0+c_m), c_m. \end{array} \tag{2.14}$$

By Theorem 2.1, we will have a continuous scaling function if the leading eigenvalue is one and the remaining eigenvalues are less than one in absolute value. Therefore, we will have a continuous scaling function if

$$c_0 + c_m = 1/2^{m-2}, \quad |c_0| < 1, \quad |c_m| < 1, \tag{2.15}$$

and the remaining c_k's are determined by (2.12). We summarize this result in the following theorem.

THEOREM 2.15. *If*

$$c_0 + c_m = 1/2^{m-2}, \quad |c_0| < 1, \quad |c_m| < 1,$$

and

$$c_k = c_0 \binom{m-1}{k} + c_m \binom{m-1}{k-1},$$

then $P(T_0, T_1, x)$ gives the normalized continuous solution of (2.1).

2.7. Analysis of smooth scaling functions. If the inequalities in (2.15) are made stricter by a factor of $1/2^\ell$, then the degree of smoothness of ϕ increases by ℓ . This is expressed in the following theorem.

THEOREM 2.16. *If*

$$c_0 + c_m = 1/2^{m-2}, \quad |c_0| < 1/2^\ell, \quad |c_m| < 1/2^\ell$$

for an integer $0 \le \ell < m-1$ *and*

$$c_k = c_0 \binom{m-1}{k} + c_m \binom{m-1}{k-1},$$

then $P(T_0, T_1, x)$ *gives the normalized* ℓ *times continuously differentiable solution of* (2.1).

Proof. This can be shown by considering the divided difference of ϕ:

$$\Delta(\phi, \ell, h, x) = \frac{1}{h^\ell} \sum_{i=0}^{\ell} (-1)^i \binom{\ell}{i} \phi(x + (\ell - i)h). \tag{2.16}$$

If the limit of above expression, as $h \to 0$, exists, then ϕ is ℓ times differentiable. We will use the matrix form of (2.16) and some of the results from Theorem A.1 (see Appendix A) to prove this theorem.

Consider the binary expansion of the numbers $x + jh$ for $0 \le j \le \ell$. If x is not a dyadic, then, as $h \to 0$, the number of common initial digits of the numbers $x + jh$ will tend to infinity. To ensure the same for dyadic x, we use the expansion of x that ends in a tail of zeros if $h > 0$; however, if $h < 0$, then we use the expansion that ends in a tail of ones. Suppose that the binary expansions of $x + jh$'s differ only on the k (possibly infinite) digits in the positions $n+1$ through $n+k$. Then we may write $x + jh = y + 2^{-n} w_j + 2^{-n-k} z$, where w_j's, for $0 \le j \le \ell$, are equidistant numbers, y represents the initial common digits, and z represents the ending common digits (if any). Here y, z, and w_j's are in the unit interval. Let $\hbar = 2^n h$, $\theta = w_0$, and define $\tilde{\mathcal{D}} = \tilde{\mathcal{D}}(k, \ell, \hbar, \theta)$, as in Theorem A.1, by

$$\tilde{\mathcal{D}} = \tilde{\mathcal{D}}(k, \ell, \hbar, \theta) = \frac{1}{\hbar^\ell} \sum_{i=0}^{\ell} (-1)^i \binom{\ell}{i} \tilde{P}_k(w_{\ell - i}). \tag{2.17}$$

Then the triangularization of the matrix form of (2.16) leads to

$$\tilde{Q}_n(x) = S^{-1} \Delta(\Phi, \ell, h, x) S = 2^{n\ell} \tilde{P}_n(x) \tilde{\mathcal{D}}(k, \ell, \hbar, \theta) \tilde{P}(z). \tag{2.18}$$

While the first ℓ columns of $2^{n\ell} \tilde{P}_n(x)$ grow unbounded as $n \to \infty$, they are nullified by the first ℓ rows of $\tilde{\mathcal{D}}$, which are zero. The $(\ell+1)$-st column of $2^{n\ell} \tilde{P}_n(x)$ is finite, and its diagonal entry is 1. This column multiplies the first entry of the $(\ell+1)$-st row of $\tilde{\mathcal{D}}$, which is $\ell!$. The remaining columns of $2^{n\ell} \tilde{P}_n(x)$ for $j > \ell + 1$ and elements of $\tilde{\mathcal{D}}$ for $i - j < \ell$ are zero. Also notice that $\tilde{P}(z)_{11} = 1$ and $\tilde{P}(z)_{ij} = 0$ for $j > 1$. As a result, all columns of $\tilde{Q} = \lim_{n \to \infty} \tilde{Q}_n$ beyond the first one are zero. Moreover, the first column is simply the "normalized" form of the $(\ell+1)$-st column of $\tilde{P}(x)$, that is,

$$\tilde{Q}(x)_{i,1} = \ell! \lim_{n \to \infty} 2^{n\ell} \tilde{P}_n(x)_{i,\ell+1}. \tag{2.19}$$

The existence of this limit follows from Remark 2.2.

To establish the continuity of the ℓth derivative, we note that the pair of eigenvectors of $\tilde{T}_0$ and $\tilde{T}_1$ corresponding to the eigenvalue $1/2^\ell$ are, by Lemma 2.7, consistent and shift continuous. Then an argument similar to Lemma 2.5 or Theorem 2.2 shows that $\tilde{Q}$ is continuous and shift continuous. Hence ϕ is ℓ times continuously differentiable. □

Remark 2.3. If ℓ is allowed to be a real number, $\ell = [\ell] + r$, $0 \le r < 1$, then $d\phi^{[\ell]}/dx^{[\ell]}$ is Hölder continuous with exponent at least r. This follows from considering the submatrices obtained by removing the first $[\ell]$ rows and columns of $2^{[\ell]}\tilde{T}_0$, $2^{[\ell]}\tilde{T}_1$, and applying the methods of Remark 2.2 and Lemma 2.6.

3. *N*-dimensional scaling functions. Higher-dimensional wavelets and scaling functions are important in analyzing multivariable cases. Rectangular wavelets can be constructed for $\mathcal{R}^N$ in ways similar to the one-dimensional case [7, Chapter 10]. However, as the number of dimensions and coefficients increases, it becomes less practical to ascertain regularity properties of the general N-dimensional scaling functions and the corresponding wavelets.

Our aim in this section is to identify a class of smooth scaling functions by generalizing the results from section 2 to N dimensions. In order to abbreviate the formulas and compare quantities in N dimensions, we first introduce a few notations. Assume $X = (X_1, X_2, \ldots, X_N)$ and $Y = (Y_1, Y_2, \ldots, Y_N)$ are two N-tuples. We define *reverse lexicographic order* $X \prec Y$ to mean $X_N < Y_N$, or there is $1 \le n < N$ so that $X_n < Y_n$ and $X_{n'} = Y_{n'}$, for $n' > n$. For the N-tuple $I \in \{1, \ldots, m\}^N$, we define $\hat{I} = 1 + \sum_{n=1}^{N} m^{n-1}(I_n - 1)$. The hat function enumerates the cells in the N-cube, $[0, m]^N$, from 1 to m^N, by going through components with lower indices first. Notice $\hat{I} < \hat{J}$ iff $I \prec J$. Finally, if s is a scalar, and it is added or compared to a vector, then s stands for $(s, s, \ldots, s)$.

Now, to generalize (2.2) to N dimensions, let $H_\mu \in \{0, \ldots, m-1\}^N$ for $\mu = 1, \ldots, m^N$ be ordered by $(0, \ldots, 0) = H_1 \prec H_2 \prec \cdots \prec H_{m^N} = (m-1, \ldots, m-1)$. Then we have

$$\Phi(X) = [\phi(X + H_1), \phi(X + H_2), \ldots, \phi(X + H_{m^N})]^t \quad \text{for} \quad X \in [0, 1]^N, \tag{3.1a}$$

$$C_K = 0 \quad \text{for} \quad K \notin \{0, \ldots, m\}^N, \tag{3.1b}$$

$$(T_D)_{\hat{I}\hat{J}} = C_{2I-J+D-1} \quad \text{for} \quad I, J \in \{1, \ldots, m\}^N \quad \text{and} \quad D \in \{0, 1\}^N. \tag{3.1c}$$

(In what follows, the range of I, J, and D is as in (3.1c) unless further restricted.) Notice that T_D is an $m^N \times m^N$ matrix. To identify a particular entry for a given I and J, first we divide the matrix into $m \times m$ subsquares and locate the square at the position (I_N, J_N); then we divide this $m^{N-1} \times m^{N-1}$ matrix into m by m subsquares and locate the square at the position (I_{N-1}, J_{N-1}) and so on until (I_1, J_1) is located in the final $m \times m$ matrix. In this manner we have N nested grids on each T_D. We label these grids as *level* N (for the coarsest) through level 1 (for the finest). The value of D_n determines the nth component of the index of C_K. Inside each level-n grid element this component is fixed, and across the grid elements its values changes in a pattern similar to the indexing of the matrix for the one-dimensional problem, i.e., T_{D_n}. The triangularization steps (see section 3.2 below) will utilize these grids. The statements and proofs of the one-dimensional case are easily generalized to the N-dimensional case by using these grids.

To generalize the iteration formula (2.3), first we define the following convention. We will show the qth digit of the nth component of X by $X_{n,q}$. Then, $X_{*,q}$ will indicate the vector of such digits. Similarly, $\bar{X}_{n,q}$ and $\bar{X}_{*,q}$ will be used to indicate the residual after the qth digit. Now, any continuous solution of (1.1) with its support in $[0,m]^N$ satisfies

$$\Phi(X) = T_{X_{*,1}}\Phi(2X - X_{*,1}) = T_{X_{*,1}}\Phi(\bar{X}_{*,1}), \tag{3.2}$$

and the repeated applications of (3.2) result in

$$\Phi(X) = \prod_{\ell=1}^{q} T_{X_{*,\ell}}\Phi(\bar{X}_{*,q}). \tag{3.3}$$

We define $P_q(\{T_D\}, X) = \prod_{\ell=1}^{q} T_{X_{*,\ell}}$, and $P(\{T_D\}, X) = \lim_{q\to\infty} P_q(\{T_D\}, X) =$ whenever the limit exits.

Suppose $X \in [0,1]^N$ and $1 \le n \le N$, and define $X\rceil n$ (respectively, $X\lfloor n$) to be a vector which is same as X except that its nth component is 1 (respectively, 0). We define $F(X) : [0,1]^N \to \mathcal{R}^{m^N}$ to be shift continuous if for any n, $1 \le n \le N$, $F(X\lfloor n)_{\hat{J}} = F(X\rceil n)_{\hat{J}}$ whenever $I - J = 0\rceil n$. (Here, the cell I is immediately after the cell J in the direction of the nth axis. The corresponding components of F are required to have the same value on the common face between the two cells.) Now, $\phi(X)$ is continuous on $[0,m]^N$ iff $\Phi(X)$ is continuous on $[0,1]^N$ and shift continuous. (As Φ is determined from its values at $\{0,1\}^N$, the requirement of shift continuity may also be limited to this set.)

The notion of consistency of matrices (as it appears in Definition 2.2 for base 2 and in Theorem 2.3 for bases larger than 2) can be generalized to higher dimensions in a componentwise fashion. For example, consider a set of matrices A_D for $D \in \{0,1\}^N$. These matrices are called consistent if for any n, $1 \le n \le N$, the pair $A_{D\lfloor n}, A_{D\rceil n}$ are consistent.

THEOREM 3.1. *Suppose that $\{A_D\}$ are jointly tied; then, for a given binary expansion of X, $P(\{A_D\}, X)$ exists. P is well defined and continuous at X if all components of X are nondyadic. If the matrices are consistent, then P is well defined and continuous at any X. P is Hölder continuous with exponent at least $-\log_2(\delta)$ if $1 > \delta > |\lambda|$, where λ is any nonleading eigenvalue of any A_D.*

Proof. The existence of P is proved in the same manner as in the one-dimensional case (Theorem 2.1). To prove continuity or obtain the Hölder exponent, we estimate $|P(Y) - P(X)|$ through the triangle inequality. Consider an N-cube with X and Y as two diagonally opposite corners. Define a set of points Z_n, $1 \le n \le p = 2^{N-2} + 2$, which start at X, go through the vertices of the N-cube, and arrive at Y. We have $|P(Y) - P(X)| \le \sum_{n=1}^{p-1} |P(Z_{n+1}) - P(Z_n)|$. Each consecutive pair of vertices differ in only one coordinate, and hence, we may apply the estimates in Lemma 2.5 or 2.6 to each term of the sum. (The same estimates cannot be applied to $P(Y) - P(X)$ directly because different components of X and Y may approach each other at different rates or some components may be dyadic while others are nondyadic.) The remaining steps in the proof are similar to the one-dimensional case. □

Suppose that Ω is a sublist of $(1,\ldots,N)$, i.e., $\Omega = (n_1, n_2, \ldots)$ and $1 \le n_1 < n_2 < \cdots \le N$. We say K is even (odd) on Ω if $K_{n_1}, K_{n_2}, \ldots$ are even (odd) and the remaining components of K are odd (even). Now the sum rule (2.7) can be generalized as follows.

DEFINITION 3.2. *We say the coefficients C_K satisfy the unit column sum rule if*

$$\sum_{K \text{ odd on } \Omega} C_K = \sum_{K \text{ even on } \Omega} C_K = 1 \quad \textit{for every } \Omega. \tag{3.4}$$

Notice that if C_K's satisfy this property, then the column sum for any column of any T_D is 1. In that case, all matrices have a left eigenvector of the form $(1, 1, \ldots, 1)$. As in the one-dimensional case, there is a matrix G, obtained by eliminating certain rows and columns of the matrices, which has a similar left eigenvector. To obtain G, we start from any T_D and for each n eliminate the first (respectively, the last) row and column of each of the level-n grid elements if D_n is zero (respectively, one). Thus G is an $(m-1)^N \times (m-1)^N$ matrix and is given by $G_{\hat{I}',\hat{J}'} = C_{2I'-J'}$, where I' and J' are taken from $\{1, \ldots, m-1\}^N$ and $\hat{I}' = 1 + \sum_{n=1}^{N} (m-1)^{n-1}(I'_n - 1)$. Any right eigenvector of G can be extended to an eigenvector of T_D by padding it with zeros at the locations where the rows of T_D were eliminated. In particular the right eigenvector corresponding to 1 generates an eigenvector for each T_D. If 1 is a simple eigenvalue, then the consistency of $\{T_D\}$ easily follows.

We are looking for the unique normalized continuous solution of (1.1) with support in $[0, m]^N$. According to (3.2) such a solution will satisfy certain simple and important properties (similar to (2.4)) as follows:

$$T_D \Phi(D) = \Phi(D), \tag{3.5a}$$

$$\Phi\left(\frac{D\lfloor n + D\rceil n}{2}\right) = T_{D\lfloor n}\Phi(D\rceil n) = T_{D\rceil n}\Phi(D\lfloor n), \tag{3.5b}$$

and in general for any $X \in [0, 1]^N$

$$\Phi\left(\frac{X\lfloor n + X\rceil n}{2}\right) = T_{X_{*,1}\rceil n}\Phi(\bar{X}_{*,1}\lfloor n) = T_{X_{*,1}\lfloor n}\Phi(\bar{X}_{*,1}\rceil n). \tag{3.5c}$$

Other properties of these solutions (similar to the "sum rules" in Lemma 2.8) are expressed as follows.

LEMMA 3.3. *Let ϕ be a continuous solution of* (1.1)*, and assume $\Gamma = \int \phi(X)dX \neq 0$. Then*

$$\sum_K C_K = 2^N. \tag{3.6}$$

If $\Gamma = 1$, then

$$\sum_{\Omega} \left(\sum_{K \text{ even on } \Omega} C_K \sum_{K \text{ odd on } \Omega} \phi(K) + \sum_{K \text{ odd on } \Omega} C_K \sum_{K \text{ even on } \Omega} \phi(K) \right) = 1. \tag{3.7}$$

Moreover, if C_K's satisfy the unit column sum rule and $\Gamma = 1$, then for all X we have

$$\sum_K \phi(X + K) = 1. \tag{3.8}$$

Proof. This is identical to the proof for the one-dimensional case. The integration of (1.1) gives (3.6). A Riemann sum approximation to the integral provides (3.7) and its special case (3.8). □

THEOREM 3.4. *If C_K's satisfy the unit column sum rule and T_D's are jointly tied, then $P(\{T_D\}, X)$ exists, is continuous, and has identical columns. The dilation equation* (1.1) *has a continuous solution given by $\phi(X + I - 1) = P(\{T_D\}, X)_{\hat{I},\hat{J}}$ for any $\hat{J}$. Moreover, this solution is properly normalized, i.e., $\int \phi(X)dX = 1$.*

Proof. The proof is identical to the proof for the one-dimensional case, Theorem 2.2. □

3.1. Analysis of 3^2-term dilation equations. In this section, we give an example based on $m = 2$ and $N = 2$. We have $\Phi(X_1, X_2) = [\phi(X_1, X_2), \phi(X_1 + 1, X_2), \phi(X_1, X_2 + 1), \phi(X_1 + 1, X_2 + 1)]^t$ and

$$T_{00} = \begin{pmatrix} C_{00} & 0 & 0 & 0 \\ C_{20} & C_{10} & 0 & 0 \\ C_{02} & 0 & C_{01} & 0 \\ C_{22} & C_{12} & C_{21} & C_{11} \end{pmatrix}, \quad T_{01} = \begin{pmatrix} C_{01} & 0 & C_{00} & 0 \\ C_{21} & C_{11} & C_{20} & C_{10} \\ 0 & 0 & C_{02} & 0 \\ 0 & 0 & C_{22} & C_{12} \end{pmatrix},$$

$$T_{10} = \begin{pmatrix} C_{10} & C_{00} & 0 & 0 \\ 0 & C_{20} & 0 & 0 \\ C_{12} & C_{02} & C_{11} & C_{01} \\ 0 & C_{22} & 0 & C_{21} \end{pmatrix}, \quad T_{11} = \begin{pmatrix} C_{11} & C_{01} & C_{10} & C_{00} \\ 0 & C_{21} & 0 & C_{20} \\ 0 & 0 & C_{12} & C_{02} \\ 0 & 0 & 0 & C_{22} \end{pmatrix}.$$

We can triangularize these matrices in two steps if

$$\begin{array}{l} C_{10} = C_{00} + C_{20},\ C_{01} = C_{00} + C_{02},\ C_{12} = C_{02} + C_{22},\ C_{21} = C_{20} + C_{22}, \\ C_{11} = C_{01} + C_{21} = C_{10} + C_{12} = C_{00} + C_{20} + C_{22} + C_{02}. \end{array} \tag{3.9}$$

Let $\mathbf{I}_n$ denote an $n \times n$ identity matrix, and define

$$S = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 0 & \mathbf{I}_2 \\ \mathbf{I}_2 & -\mathbf{I}_2 \end{pmatrix}, \quad S_1 = \begin{pmatrix} S & 0 \\ 0 & S \end{pmatrix},$$

$$S^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad {S_2}^{-1} = \begin{pmatrix} \mathbf{I}_2 & \mathbf{I}_2 \\ \mathbf{I}_2 & 0 \end{pmatrix}, \quad {S_1}^{-1} = \begin{pmatrix} S^{-1} & 0 \\ 0 & S^{-1} \end{pmatrix}.$$

Then, any of matrices T_D, $D \in \{0, 1\}^2$, can be triangularized by

$$\tilde{T}_D = {S_1}^{-1}{S_2}^{-1} T_D S_2 S_1.$$

The four eigenvalues of each of the four matrices may be given in terms of the "corner" C_K's as

$$\begin{array}{l} (C_{00} + C_{20} + C_{22} + C_{02},\ C_{00} + C_{02},\ C_{00} + C_{20},\ C_{00}), \\ (C_{00} + C_{20} + C_{22} + C_{02},\ C_{00} + C_{02},\ C_{02} + C_{22},\ C_{02}), \\ (C_{00} + C_{20} + C_{22} + C_{02},\ C_{22} + C_{20},\ C_{00} + C_{20},\ C_{20}), \\ (C_{00} + C_{20} + C_{22} + C_{02},\ C_{22} + C_{20},\ C_{02} + C_{22},\ C_{22}). \end{array}$$

Notice that sums of C_K's on every corner, corners of every side, and corners of the entire square show up as eigenvalues. For convergence of the products of T_D's, we need the leading eigenvalue to be 1 and the remaining eigenvalues to be less than 1

in absolute value. The result can be displayed in terms of a three-parameter space, say, (C_{00}, C_{02}, C_{20}). Therefore, we require $C_{22} = 1 - C_{00} - C_{02} - C_{20}$ and

$$-1 < C_{00} < 1,\ -1 < C_{02} < 1,\ -1 < C_{20} < 1,$$
$$0 < C_{00} + C_{02} < 1,\ 0 < C_{00} + C_{20} < 1,\ 0 < C_{00} + C_{02} + C_{20} < 2.$$

The remaining C_K's are then determined by (3.9). Notice that this solution has three degrees of freedom and can change its sign on $[0, 2]^2$, while a tensor product solution $\phi(x, y) = \phi_1(x)\phi_2(y)$, where ϕ_1 and ϕ_2 satisfy

$$\phi_1(x) = \alpha\phi_1(2x) + \phi_1(2x - 1) + (1 - \alpha)\phi_1(2x - 2),$$
$$\phi_2(y) = \beta\phi_2(2y) + \phi_2(2y - 1) + (1 - \beta)\phi_2(2y - 2),$$

has only two degrees of freedom, i.e., $0 < \alpha, \beta < 1$, and a fixed sign.

3.2. Analysis of $(m + 1)^N$-term dilation equations. In this section, we consider the dilation equations for a given m and N, with coefficients that satisfy (2.11) along every coordinate direction. In this case the coefficients, C_K, are given by binomial interpolation of their values on the corners of the N-cube, that is, C_{mD}'s. Applying (2.12) repeatedly along all coordinate directions gives

$$C_K = \sum_{D \in \{0,1\}^N} C_{mD} \prod_{n=1}^{N} \binom{m-1}{K_n - D_n}. \tag{3.10}$$

Now, all 2^N corresponding T_D's can be triangularized by a set of matrices built from S given by (2.13). These matrices are constructed as follows: given an n, $1 \le n \le N$, first replace every entry S_{ij} of S with a diagonal matrix $S_{ij}\mathbf{I}_{m^n}$. This will produce an $m^{n+1} \times m^{n+1}$ intermediate matrix. Then, use m^{N-n-1} copies of the intermediate matrix to create S_n, a block diagonal matrix of size $m^N \times m^N$. Now, we have simultaneous triangularization by

$$\tilde{T}_D = {S_1}^{-1}{S_2}^{-1} \cdots {S_{N-1}}^{-1}{S_N}^{-1} T_D S_N S_{N-1} \cdots S_2 S_1.$$

Here, at each stage n, $n = N, \ldots, 1$, the effect of S_n and ${S_n}^{-1}$ is to triangularize the current level-n grid elements. The formation of the resulting diagonal blocks is similar to the one-dimensional formula (2.14). The entries that act as c_0 and c_m are a pair of blocks, within each grid's $m \times m$ subdivision, in the positions $(1, 1 + D_n)$ and $(m, m + D_n - 1)$. We call these the *polar* blocks. At the end of each stage of triangularization the polar blocks and their sum, scaled by factors $1, \ldots, 2^{m-2}$, appear on the diagonal. The first entry of T_D is C_D, and the corresponding indices of C_K's in each pair of polar blocks differ in only one component. As a result, the final summation of C_K's occur on the corners of the faces of the N-cube. If the face is n-dimensional, then the sum will appear with scale factors of up to $2^{(m-2)n}$.

To formally describe the eigenvalues of T_D's, i.e., the diagonal elements of $\tilde{T}_D$'s, we need to construct sums of C_K's on the corners of every n-dimensional face of the N-cube, $\{0, m\}^N$. Let θ be a sublist of $(1, \ldots, N)$, $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, $1 \le \theta_1 < \theta_2 < \cdots < \theta_n \le N$, and assume that θ' is its complimentary list, $\theta \cup \theta' = \{1, \ldots, N\}$. (If θ is empty, then $n = 0$.) Now, let $D_\theta = (D_{\theta_1}, D_{\theta_2}, \ldots, D_{\theta_n})$ be a free element of $\{0, 1\}^n$, and assume that the remaining elements of D form a fixed element of $\{0, 1\}^{N-n}$, say, $D_{\theta'} = \Upsilon$. Then, for every Υ there are diagonal entries (with various multiplicities) of

$\tilde{T}_D$ in the form

$$2^{\gamma} \sum_{\substack{D_\theta \\ D_{\theta'}=\Upsilon}} C_{mD} \quad \text{for} \quad 0 \le \gamma \le (m-2)n.$$

Hence, for the convergence of the matrix products to a continuous function, we require

$$\sum_{D} C_{mD} = 1/2^{(m-2)N}, \tag{3.11}$$

which restricts the sum of C_K's on all corners of the N-cube. Similarly, for every n-dimensional ($n < N$) face of the cube, described by an Υ, we require

$$\left| \sum_{\substack{D_\theta \\ D_{\theta'}=\Upsilon}} C_{mD} \right| < 1/2^{(m-2)n}. \tag{3.12}$$

Now, (3.10)–(3.12) characterize a class of C_K's which produce continuous scaling functions. As (3.11) indicates, there are $2^N - 1$ degrees of freedom. (For each $n < N$ there are $\binom{N}{n}2^{N-n}$ inequalities of the form (3.12). There are $3^N - 1$ inequalities in all.)

As in the case of the one-dimensional scaling functions, one can increase the degree of smoothness of ϕ by ℓ if the inequalities (3.12) are made stricter by a factor of $1/2^{\ell}$. We summarize our results in the following theorem.

THEOREM 3.5. *If the sum of C_K's on the corners of the N-cube satisfies*

$$\sum_{D} C_{mD} = 1/2^{(m-2)N},$$

and for every n-dimensional ($n < N$) face of the N-cube we have

$$\left| \sum_{\substack{D_\theta \\ D_{\theta'}=\Upsilon}} C_{mD} \right| < 1/2^{(m-2)n+\ell},$$

and all other C_K's are given by binomial interpolation of their values at the corners of the N-cube,

$$C_K = \sum_{D \in \{0,1\}^N} C_{mD} \prod_{n=1}^{N} \binom{m-1}{K_n - D_n},$$

then the solution of (1.1) *is ℓ times continuously differentiable. If ℓ is allowed to be a real number, then the $[\ell]$th derivative of ϕ is Hölder continuous with exponent at least $\ell - [\ell]$.*

Proof. If $\ell = 0$, then we establish continuity by using Theorem 3.2. For $\ell > 0$ we investigate existence, continuity, and the Hölder exponent of the required derivative of ϕ by considering the related partial derivatives. The treatment is analogous to the one-dimensional case, and it uses the generalization of Theorem A.1 to N dimensions. □

3.3. Examples of 4^2-term scaling functions. In this section we give some pictorial examples of the scaling functions obtained by applying (3.10)–(3.12) to the case of $m = 3$ and $N = 2$. We require

$$C_{00} + C_{03} + C_{30} + C_{33} = 1/2^2,$$
$$|C_{00}+C_{03}| < 1/2^{1+\ell},\ |C_{03}+C_{33}| < 1/2^{1+\ell},\ |C_{33}+C_{30}| < 1/2^{1+\ell},\ |C_{30}+C_{00}| < 1/2^{1+\ell},$$
$$|C_{00}| < 1/2^{\ell},\ |C_{03}| < 1/2^{\ell},\ |C_{30}| < 1/2^{\ell},\ |C_{33}| < 1/2^{\ell}.$$

The remaining C_K's are given by (3.10), which in two dimensions reads

$$C_{ij} = C_{00}\binom{m-1}{i}\binom{m-1}{j} + C_{0m}\binom{m-1}{i}\binom{m-1}{j-1}$$
$$+C_{m0}\binom{m-1}{i-1}\binom{m-1}{j} + C_{mm}\binom{m-1}{i-1}\binom{m-1}{j-1}.$$

In our example, $m = 3$, $(i, j) \in \{0, 1, 2, 3\}^2$, and the support of ϕ is $[0, 3]^2$. For a given set of coefficients, C_K, the supremum of all possible real values of ℓ will be shown by ℓ_c, the critical exponent. Then, for any $\ell < \ell_c$, the $[\ell]$th derivative of ϕ is Hölder continuous with exponent (at least) $\ell - [\ell]$.

If we choose $C_{00} = C_{03} = C_{30} = C_{33} = 1/16$, then we get the two-dimensional spline in Figure 1. This function fails to have continuous second derivatives at points in its support where one of the coordinates is an integer. The maximum integer value that we can use for ℓ in Theorem 3.3 is 1. Therefore, this function is $\mathcal{C}^1$. The critical exponent is $\ell_c = 2$.

If we choose $C_{00} = -0.075$, $C_{03} = C_{30} = 0.1$, and $C_{33} = 0.125$, then we get the graph in Figure 2. The maximum value that we can use for ℓ in Theorem 3.3 is 1. Therefore, this function is $\mathcal{C}^1$, despite appearances. The critical exponent is $\ell_c = -\log_2 0.45 = 1.152\ldots$.

If we choose $C_{00} = -0.5$, $C_{03} = C_{30} = 0.625$, and $C_{33} = -0.5$, then we get the graph in Figure 3. The maximum value that we can use for ℓ in Theorem 3.3 is 0. Therefore this function is only continuous. The critical exponent is $\ell_c = -\log_2 0.625 = 0.678\ldots$.

Appendix A. Some notes on binomial identities. Here we outline the proofs of some binomial identities used in this paper.

Note A.1. To verify that (2.12) is a solution of (2.11), one evaluates

$$\left(x\frac{d}{dx}\right)^q [x^b(1-x)^{m-1}] = \sum_{k=0}^{m}(-1)^k(k+b)^q\binom{m-1}{k}x^{k+b} \tag{A.1}$$

for $0 \le q \le m-2$, and $b = 0$ or -1 at $x = 1$.

The fact that matrices given by (2.13) are inverses of each other follows from (A.1) for $q = b = 0$.

Note A.2. The matrices $\tilde{T}_0$ and $\tilde{T}_1$ and their divided differences have a particular zero structure and a simple formula for the entries of the first nonzero subdiagonal. The divided differences in question are polynomials such as $\tilde{T}_1 - \tilde{T}_0$, $\tilde{T}_1\tilde{T}_1 - 2\tilde{T}_1\tilde{T}_0 + \tilde{T}_0\tilde{T}_1$, etc., where the indices form an arithmetic sequence of binary numbers and the coefficients are the binomial numbers. This is discussed in the following theorem.

THEOREM A.1. *Consider a set of equidistant numbers $w_j = \theta + j\hbar$ for $j = 0, \ldots, \ell$, in the unit interval and with binary expansions that differ on the first k digits only.*

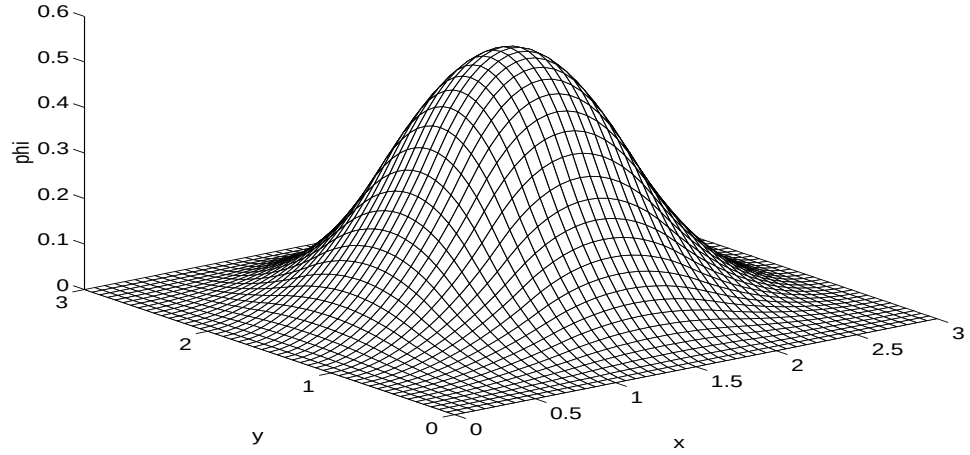


FIG. 3.1. *The scaling function for* $C_{00} = C_{03} = C_{30} = C_{33} = 1/16$. *Here,* $\ell_c = 2$.

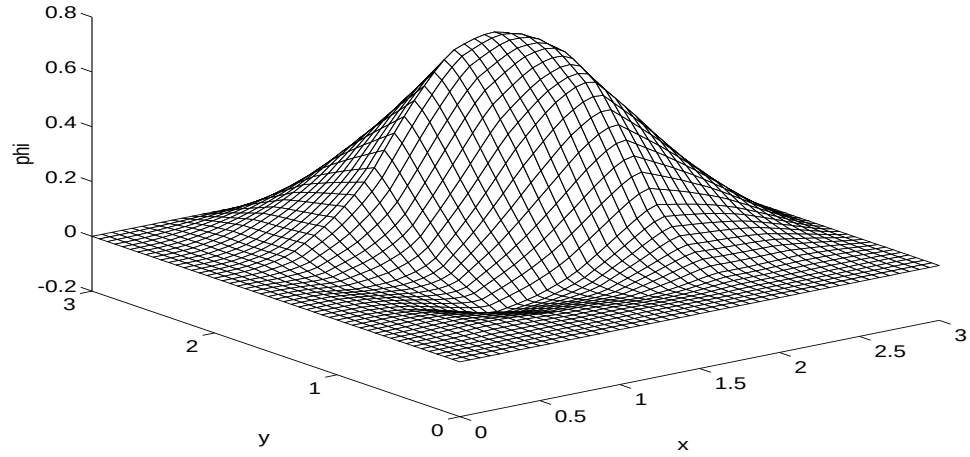


FIG. 3.2. *The scaling function for* $C_{00} = -0.075$, $C_{03} = C_{30} = 0.1$, $C_{33} = 0.125$. *Here,* $\ell_c = 1.152\ldots$.

Define

$$\tilde{\mathcal{D}} = \tilde{\mathcal{D}}(k, \ell, \hbar, \theta) = \frac{1}{\hbar^\ell} \sum_{i=0}^{\ell} (-1)^i \binom{\ell}{i} \tilde{P}_k(w_{\ell - i}). \tag{A.2}$$

Then

$$\tilde{\mathcal{D}}_{i,j} = 0 \quad \textit{if} \quad i < m \quad \textit{and} \quad i < j + \ell, \tag{A.3}$$

$$\tilde{\mathcal{D}}_{m,j} = 0 \quad \textit{if} \quad c_0 = c_m \quad \textit{and} \quad m < j + \ell, \tag{A.4}$$

$$\tilde{\mathcal{D}}_{i,i-\ell} = \ell! \binom{i-1}{\ell} 2^{k(m+\ell-i-1)} (c_0 + c_m)^k \quad \textit{for} \quad \ell < i < m. \tag{A.5}$$

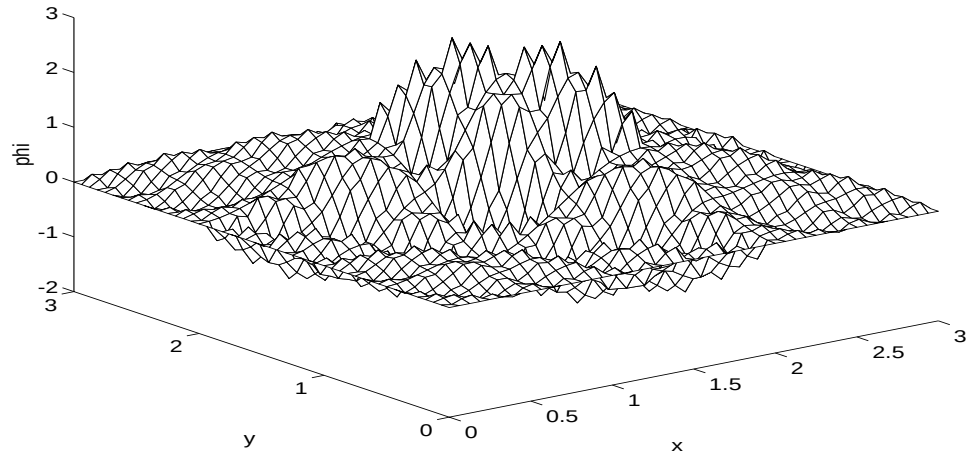


FIG. 3.3. *The scaling function for* $C_{00} = -0.5$, $C_{03} = C_{30} = 0.625$, *and* $C_{33} = -0.5$. *Here,* $\ell_c = 0.678\ldots$.

In particular, if c_k*'s satisfy the unit column sum rule, then* $c_0 + c_k = 1/2^{m-2}$ *and we have*

$$\tilde{\mathcal{D}}_{i,i-\ell} = \ell! \binom{i-1}{\ell} 2^{k(\ell+1-i)}. \tag{A.6}$$

The first entry of this list is

$$\tilde{\mathcal{D}}_{\ell+1,1} = \ell!. \tag{A.7}$$

Proof. We prove the theorem for $\tilde{T}_1$ and $\tilde{T}_0$. The general case is similar and can be shown by induction on ℓ. The proof rests on simple divided difference properties of polynomials. We establish that $S^{-1}T_0S$ and $S^{-1}T_1S$ are lower triangular, with diagonal entries given by (2.14). Define the matrices M_a for $a = 0, 1, 2$ by $(M_a)_{i,j} = \binom{m-1}{2i-j-a}$. We have $T_0 = c_0M_1 + c_mM_2$ and $T_1 = c_0M_0 + c_mM_1$. First, we show that $S^{-1}M_aS$ is lower triangular and determine its diagonal entries. The main step is to prove the following identity:

$$\sum_{1\le k,l\le m} (i-1)! \binom{m-k}{i-1} \binom{m-1}{2k-l-a} \binom{j-1}{m-l} \frac{(-1)^{l+j-m-1}}{(j-1)!} = 0 \tag{A.8}$$

for $j > i$ and $a = 0, 1, 2$.

A brief outline of the proof of (A.8) is as follows. First, we notice that for each fixed i the elements of row i of S^{-1} are the values of a polynomial of order $i-1$ in j. Then we will show that the same is true of $S^{-1}M_a$ (except that when $a = 0$, the last row is a polynomial of order $m-2$, and when $a = 2$, the last row is zero). Next we observe that for each j the elements of the column j of S are proportional to the coefficients of the divided difference scheme of order $j-1$. Therefore, when $j > i$, the product of row i of $S^{-1}M_a$ and column j of S is zero; hence, (A.8) follows.

Now we show that row i of $S^{-1}M_a$ is a polynomial of degree at most $i-1$. For a fixed i let $s(j) = S^{-1}_{ij}$. Obviously s is polynomial of degree $i-1$. Define g and h in

terms of alternate values of $S^{-1}M_a$ on row i; that is,

$$\sum_{1\le k\le m}(i-1)!\binom{m-k}{i-1}\binom{m-1}{2k-l-a}=\begin{cases}g(l) & \text{for } l \text{ odd},\\ h(l) & \text{for } l \text{ even}.\end{cases} \tag{A.9}$$

The alternate columns of S_{ij}^{-1} are identical up to a shift. Hence, for the j's with same parity, we get g as the same linear combination of translates of s. Therefore, g is a polynomial of degree $i-1$. A similar argument applies to h. We will show that for $i<m$ these two polynomials are identical. Define $f(x)=s(x/2)$, and notice that the leading term of f is $(-x/2)^{i-1}$. Assume $a=0$ (the cases for $a=1$ and $a=2$ can be treated similarly). Then we write (A.9) as

$$\begin{aligned} g(x)&=\textstyle\sum_r\binom{m-1}{2r-1}f(x+2r-1),\\ h(x)&=\textstyle\sum_r\binom{m-1}{2r}f(x+2r).\end{aligned} \tag{A.10}$$

Suppose $f(x)=\sum_p a_p x^p$, where $a_p=0$ for $p<0$ or $p>i-1$. Then from binomial expansion of (A.10) we obtain

$$\begin{aligned} g(x)&=\textstyle\sum_{p,q}a_p\binom{p}{q}x^{p-q}\sum_r\binom{m-1}{2r-1}(2r-1)^q,\\ h(x)&=\textstyle\sum_{p,q}a_p\binom{p}{q}x^{p-q}\sum_r\binom{m-1}{2r}(2r)^q.\end{aligned} \tag{A.11}$$

But from Note A.1 we have

$$\sum_r\binom{m-1}{2r-1}(2r-1)^q=\sum_r\binom{m-1}{2r}(2r)^q \qquad \text{for} \qquad 0\le q\le m-2; \tag{A.12}$$

therefore, $g(x)$ and $h(x)$ are identical if their degree $i-1$ does not exceed $m-2$; that is, if $i<m$. In this case the leading term of the polynomial, from (2.19), is $2^{m-2}(-x/2)^{i-1}$. Now the entries on column j of S may be written as $[(-1)^{i+j}\binom{j-1}{m-i}]\times[(-1)^{m+1}/(j-1)!]$. The first part is the divided difference scheme of order $j-1$, and the second part is a constant. Therefore, the product of row i of $S^{-1}M_a$ and column j of S is zero when $i<j$. Hence, $S^{-1}M_aS$ is lower triangular. When $i=j<m$, then the product is 2^{m-i-1}. For $i=j=m$ we need to distinguish among three cases. When $a=0$, the last row of $S^{-1}M_a$ is $[m-1,1,0,0,\dots,0]$. The interpolating polynomial of these values is of degree $m-1$. Hence, its product with column m of S is zero. When $a=2$, then the last row itself is zero. When $a=1$, we get a nonzero contribution, i.e., 1. This explains the particular form of the eigenvalues in (2.14). ∎

Acknowledgments. I am grateful to Professor Gilbert Strang for interesting me in this topic, to the referees for pointing out several corrections, and to Professor Christopher Heil for valuable suggestions.

REFERENCES

[1] M. A. Berger and Y. Wang, *Bounded semi-groups of matrices*, Linear Algebra Appl., 166 (1992), pp. 21–27.

[2] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary subdivision*, Mem. Amer. Math. Soc., 93 (1991), pp. 1–186.

[3] J. E. Cohen, *Subadditivity, generalized products of random matrices and operations research*, SIAM Rev., 30 (1988), pp. 69–86.

[4] D. Colella and C. Heil, *The characterization of continuous, four-coefficient scaling functions and wavelets*, IEEE Trans. Inform. Theory, Special Issue on Wavelet Transforms and Multiresolution Signal Analysis, 38 (1992), pp. 876–881.

[5] D. Colella and C. Heil, *Characterizations of scaling functions, continuous solutions*, SIAM J. Matrix Anal. Appl., 15 (1994), no. 2, pp. 496–518.

[6] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 49 (1988), pp. 909–996.

[7] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[8] I. Daubechies and J. Lagarias, *Two-scale difference equations,* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.

[9] I. Daubechies and J. Lagarias, *Two-scale difference equations,* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.

[10] I. Daubechies and J. Lagarias, *Sets of matrices all infinite products of which converge*, Linear Algebra Appl., 161 (1992), pp. 227–263.

[11] C. Heil and G. Strang, *Continuity of the joint spectral radius: Application to wavelets*, in Linear Algebra for Signal Processing, IMA Vol. Math. Appl. 69, A. Bojanczyk and G. Cybenko, eds., Springer-Verlag, New York, 1995. pp. 51–61,

[12] J. C. Lagarias and Y. Wang, *The finiteness conjecture for the generalized spectral radius of a set of matrices*, Linear Algebra Appl., 214 (1995), pp. 17–42.

[13] C. A. Micchelli and H. Prautzsch, *Refinement and subdivision for spaces of integer translates of a compactly supported function*, in Numerical Analysis 1987, Pitman Res. Notes Math. Ser. 170, D. F. Griffiths and G. A. Watson, eds., Longman Sci. Tech., Harlow, UK, 1988, pp. 192–222.

[14] C. A. Micchelli and H. Prautzsch, *Uniform refinement of curves,* Linear Algebra Appl., 114/115 (1989), pp. 841–870.

[15] M. Maesumi, *Optimal unit ball for joint spectral radius, an example from four-coefficient MRA*, in Approximation Theory VIII, Vol. 2: Wavelets and Multilevel Approximation, Proceedings of the Eighth International Conference on Approximation Theory, C. K. Chui, and L. L. Schumaker, eds, World Scientific, Singapore, pp. 267–274.

[16] G. C. Rota and G. Strang, *A note on the joint spectral radius*, Konink. Nederl. Akad. Wetensch. Proc. A, 63 (1960), pp. 379–381.

[17] G. Strang, *Wavelets and dilation equations: A brief introduction*, SIAM Rev., 31 (1989), pp. 614–627.

[18] G. Strang, *Wavelet transforms versus Fourier transforms*, Bull. Amer. Math. Soc. (N.S.), 28 (1993), pp. 288–305.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 1027–1039, October 1998

012

A NUMERICAL METHOD FOR THE INVERSE STOCHASTIC SPECTRUM PROBLEM*

MOODY T. CHU† AND QUANLIN GUO†

Abstract. The inverse stochastic spectrum problem involves the construction of a stochastic matrix with a prescribed spectrum. The problem could be solved by first constructing a nonnegative matrix with the same prescribed spectrum. A differential equation aimed to bring forth the steepest descent flow in reducing the distance between isospectral matrices and nonnegative matrices, represented in terms of some general coordinates, is described. The flow is further characterized by an analytic singular value decomposition to maintain the numerical stability and to monitor the proximity to singularity. This flow approach can be used to design Markov chains with specified structure. Applications are demonstrated by numerical examples.

Key words. nonnegative matrix, stochastic matrix, least squares, steepest descent, isospectral flow, structured Markov chain, analytic singular value flow

AMS subject classifications. 65F15, 65H15

PII. S0895479896292418

1. Introduction.

Inverse eigenvalue problems concern the reconstruction of matrices from prescribed spectral data. The spectral data may involve complete or partial information of eigenvalues or eigenvectors. Generally, a problem without any restrictions on the matrix is of little interest. In order for the inverse eigenvalue problem to be meaningful, it is often necessary to restrict the construction to special classes of matrices, such as symmetric Toeplitz matrices or matrices with other special structures. In this paper we limit our attention to the so-called stochastic matrices, i.e., matrices with nonnegative elements where all their row sums are equal to one. We propose a numerical procedure for the construction of a stochastic matrix so that its spectrum agrees with a prescribed set of complex values. If the set of prescribed values turns out to be infeasible, the method produces a best approximation in the sense of least squares. To our knowledge, this inverse eigenvalue problem for stochastic matrices has not been studied extensively, probably due to its difficulty as we shall discuss below. Nevertheless, for a variety of physical problems that can be described in the context of Markov chains, an understanding of the inverse eigenvalue problem for stochastic matrices and a capacity to solve the problem would make it possible to construct a system from its natural frequencies [8, 12]. The method proposed in this paper appears to be the first attempt at tackling this problem numerically with some success. Our technique can also be applied as a numerical way to solve the long standing inverse eigenvalue problems for nonnegative matrices.

Associated with every inverse eigenvalue problem are two fundamental questions: the theoretic issue on solvability and the practical issue on computability. The major effort in solvability has been to determine a necessary or sufficient condition under which an inverse eigenvalue problem has a solution, whereas the main concern in computability has been to develop an algorithm by which, knowing a priori that the

*Received by the editors March 4, 1996; accepted for publication (in revised form) by G. P. Styan July 7, 1997; published electronically July 17, 1998.
http://www.siam.org/journals/simax/19-4/29241.html
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (chu@math.ncsu.edu). The first author was supported in part by National Science Foundation grant DMS-9422280.

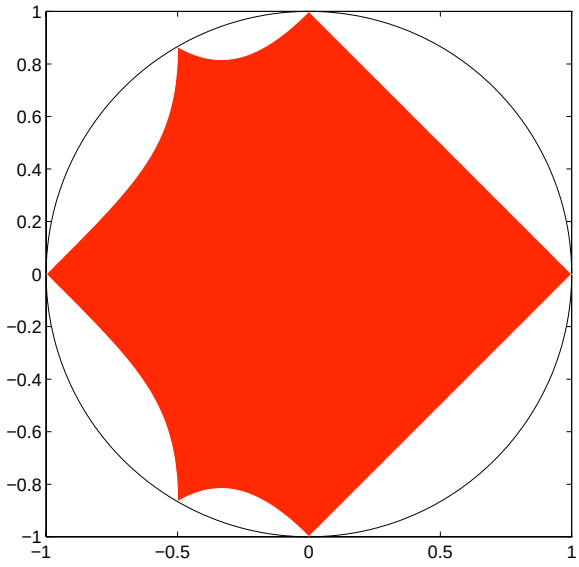


FIG. 1. Θ_4 *by the Karpelevič theorem.*

given spectral data are feasible, a matrix can be constructed numerically. Both questions are difficult and challenging. Searching through the literature, we have found only a handful of inverse eigenvalue problems that have been completely understood or solved. A collection of inverse eigenvalue problems and their status reviews can be found in a recent article [7]. The focus of this paper is on the computability for stochastic matrices.

For stochastic matrices, the inverse eigenvalue problem is particularly difficult as can be seen from the involvement in the best known result on existence by Karpelevič [15, 17]. Karpelevič completely characterized the set Θ_n of points in the complex plane that are eigenvalues of stochastic $n \times n$ matrices. In particular, the region Θ_n is symmetric about the real axis. It is contained within the unit circle and its intersections with the unit circle are points $z = e^{(2\pi a/b)i}$, where a and b run over all integers satisfying $0 \le a < b \le n$. The boundary of Θ_n consists of these intersection points and of curvilinear arcs connecting them in circular order. These arcs are characterized by specific parametric equations whose formulas can be found in [15, 17]. For example, a complex number λ is an eigenvalue for a 4×4 stochastic matrix if and only if it belongs to a region Θ_4 such as the one shown in Figure 1. Complicated though it may seem, the Karpelevič theorem characterizes only one complex value at a time and does not provide further insight into when two or more points in Θ_n are eigenvalues of the *same* stochastic matrix. Minc [17] distinctively called the problem we are considering, where the entire spectrum is given, *the inverse spectrum problem.*

It is known that the inverse eigenvalue problem for nonnegative matrices is virtually equivalent to that for stochastic matrices. For example, a complex nonzero number α is an eigenvalue of a nonnegative matrix with a positive maximal eigenvalue r if and only if α/r is an eigenvalue of a stochastic matrix. Our problem is much more complicated because it involves the entire spectrum. Fortunately, based on the following theorem we can proceed with our computation once a nonnegative matrix is found.

THEOREM 1.1. *If A is a nonnegative matrix with positive maximal eigenvalue r and a positive maximal eigenvector x, then $D^{-1}r^{-1}AD$ is a stochastic matrix where $D := \text{diag}\{x_1, \ldots, x_n\}$.*

We thus should turn our attention to the inverse eigenvalue (or spectrum) problems for nonnegative matrices, a subject that has received considerable interest in the literature. Some necessary and a few sufficient conditions on whether a given set of complex numbers could be the spectrum of a nonnegative matrix can be found, for example, in [1, 3, 9, 10, 11, 13, 14, 18, 21] and the references contained therein. Yet numerical methods for constructing such a matrix, even if the spectrum is feasible, still need to be developed. Some discussion can be found in [6, 21]. Regardless of all the efforts, the inverse eigenvalue problem for nonnegative matrices has not been completely resolved to this date.

In an earlier paper [6] the first author developed an algorithm that can construct symmetric nonnegative matrices with prescribed spectra by means of differential equations. Symmetry was needed there because the techniques by then were for flows in the group of orthogonal matrices only. Upon realizing the existence of an analytic singular value decomposition (ASVD) for a real analytic path of matrices [5, 16, 22], we are able to advance the techniques in [6] to general matrices in this paper.

This paper is organized as follows. We reformulate the inverse stochastic spectrum problem as that of finding the shortest distance between isospectral matrices and nonnegative matrices. In section 2 we introduce a general coordinate system to describe these two types of matrices and discuss how this setting naturally leads to a steepest descent flow. This approach generalizes what has been done before, but requires the inversion of matrices that is potentially dangerous. In section 3 we argue that the steepest descent flow is in fact analytic and hence an ASVD exists. We therefore are able to describe the flow by a more stable vector field. We illustrate the application of this differential equation to the inverse spectrum problem by numerical examples in section 4.

2. Basic formulation. The given spectrum $\{\lambda_1, \ldots, \lambda_n\}$ may be complex valued. It is not difficult to create a simple, say tridiagonal, real-valued matrix Λ carrying the same spectrum. For multiple eigenvalues, one should also consider the possible real-valued Jordan canonical form, depending on the geometric multiplicity. Matrices in the set

$$\mathcal{M}(\Lambda) := \{P\Lambda P^{-1} | P \in R^{n\times n} \text{ is nonsingular}\} \tag{1}$$

obviously are isospectral to Λ. Let

$$\pi(R_+^n) := \{B \circ B | B \in R^{n\times n}\} \tag{2}$$

denote the cone of all nonnegative matrices, where $A \circ B := [a_{ij}b_{ij}]$ represents the Hadamard product of matrices if $A = [a_{ij}]$ and $B = [b_{ij}]$. Our basic idea is to find the intersection of $\mathcal{M}(\Lambda)$ and $\pi(R_+^n)$. Such an intersection, if it exists, results in a nonnegative matrix isospectral to Λ. Furthermore, if the condition in Theorem 1.1 holds, i.e., if the eigenvector corresponding to the positive maximal eigenvalue is positive, then we will have solved the inverse spectrum problem for stochastic matrices by a diagonal similarity transformation. The difficulty, as we pointed out earlier, is the lack of means to determine if the given spectrum is feasible. An arbitrarily given set of values $\lambda_1, \ldots, \lambda_n$, even if $\lambda_i \in \Theta_n$ for all i, may not be the spectrum of any nonnegative matrix. In this case, it is reasonable to ask for only the best possible

approximation. To handle both problems at the same time, we reformulate the inverse spectrum problem as that of finding the shortest distance between $\mathcal{M}(\Lambda)$ and $\pi(R_+^n)$:

$$\text{minimize } F(P,R) := \frac{1}{2}\|P\Lambda P^{-1} - R \circ R\|^2, \tag{3}$$

where $\|\cdot\|$ represents the Frobenius matrix norm. Obviously, if Λ is feasible, then $F(P,R) = 0$ for some suitable P and R. Note that the variable P in (3) resides in the open set of nonsingular matrices, whereas R is simply a general matrix in $R^{n\times n}$. The optimization in (3) subjects to no other significant constraint. Since the optimization is over an unbounded open domain, it is possible that the minimum does not exist. We shall comment more on this point later.

The Fréchet derivative of F at (P,R) acting on (H,K) is calculated as follows:

$$\begin{aligned} F'(P,R)(H,K) &= \langle P\Lambda P^{-1} - R\circ R, H\Lambda P^{-1} - P\Lambda(P^{-1}HP^{-1}) - K\circ R - R\circ K\rangle \\ &= \langle (P\Lambda P^{-1} - R\circ R)P^{-T}\Lambda^T - P^{-T}\Lambda^T P^T(P\Lambda P^{-1} - R\circ R)P^{-T}, H\rangle \\ &\quad -\langle 2(P\Lambda P^{-1} - R\circ R)\circ R, K\rangle, \end{aligned} \tag{4}$$

where $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product of two matrices. Define, for abbreviation,

$$M(P) := P\Lambda P^{-1}, \tag{5}$$

$$\Delta(P,R) := M(P) - R\circ R. \tag{6}$$

The norm of $\Delta(P,R)$ represents how close we are able to solve the inverse spectrum problem. With respect to the product topology on $R^{n\times n}\times R^{n\times n}$, we can easily read off the gradient ∇F of the objective function F from (4):

$$\nabla F(P,R) = \left((\Delta(P,R)M(P)^T - M(P)^T\Delta(P,R))P^{-T}, -2\Delta(P,R)\circ R\right). \tag{7}$$

Therefore, the flow $(P(t),R(t))$ defined by the differential equations

$$\frac{dP}{dt} = [M(P)^T, \Delta(P,R)]P^{-T}, \tag{8}$$

$$\frac{dR}{dt} = 2\Delta(P,R)\circ R, \tag{9}$$

where $[\cdot,\cdot]$ denotes the Lie bracket of two matrices, signifies, in fact, the steepest descent flow for the objective function F.

An important advance we have made here is that the gradient $\nabla F(P,R)$ no longer needs to be projected as was required in [6] since P need not be orthogonal. On the other hand, a possible frailty of this advance is that the solution flow $P(t)$ is susceptible to becoming unbounded.

The differential system (8) and (9) has another interesting property that is useful for constructing Markov chains with designated structure. The Hadamard product in (9) implies that if $r_{ij} = 0$, then $\frac{dr_{ij}}{dt} = 0$. Thus the zero structure in the original matrix $R(0)$ is preserved throughout the integration. We may use this property to explore the possibility of constructing a Markov chain with prescribed linkages and spectrum.

3. ASVD flow. A somewhat worrisome feature of the differential system (8) and (9) is the involvement of P^{-1}. In this section we propose using the ASVD as a stable way to carry out the computation. Also, we pointed out earlier that the minimization (3) over two open sets may not have a minimum. It is possible during the integration that the flow $P(t)$ from one particular starting value gradually moves toward the boundary, i.e., the closed subset of singular matrices in $R^{n\times n}$, and becomes more and more nearly singular. The ASVD technique allows us to monitor the situation. If the singular values indicate that $P(t)$ is nearly rank deficient, we can abort the integration and restart from a new initial value.

An analytic singular value decomposition of the path of matrices $P(t)$ is an analytic path of factorizations

$$P(t) = X(t)S(t)Y(t)^T, \tag{10}$$

where $X(t)$ and $Y(t)$ are orthogonal and $S(t)$ is diagonal. In [5] Bunse-Gerstner et al. prove that an ASVD exists if $P(t)$ is analytic. The fact that $P(t)$ defined by (8) and (9) is indeed analytic follows from the Cauchy–Kovalevskaya theorem [19] since the coefficients of the vector field in (8) and (9) are analytic. With this understanding, we may proceed to describe the differential equations for the ASVD of $P(t)$.

It is worthy to point out that the two matrices P and R are used, respectively, as *coordinates* to describe the isospectral matrices and nonnegative matrices. We may have used more dimensions of variables than necessary to describe the underlying matrices, but that does no harm. When flows $P(t)$ and $R(t)$ are introduced, correspondingly, a flow in $\mathcal{M}(\Lambda)$ and a flow in $\pi(R_+^n)$ are also introduced. To stabilize the computation, we further describe the motion of the coordinate P by three other variables X, S, and Y according to (10). The flows of $X(t)$, $S(t)$, and $Y(t)$ can be found in the following way due to Wright [16, 22].

Differentiating both sides of (10), we obtain the following equation after some suitable multiplications:

$$X^T\frac{dP}{dt}Y = X^T\frac{dX}{dt}S + \frac{dS}{dt} + S\frac{dY^T}{dt}Y. \tag{11}$$

Define

$$Q(t) := X^T\frac{dP}{dt}Y, \tag{12}$$

$$Z(t) := X^T\frac{dX}{dt}, \tag{13}$$

$$W(t) := \frac{dY^T}{dt}Y. \tag{14}$$

Note that $Q(t)$ is known from (8), where the inverse of $P(t)$ is calculated from

$$P^{-1} = YS^{-1}X^T. \tag{15}$$

The diagonal entries of $S = \operatorname{diag}\{s_1, \ldots, s_n\}$ provide us with information about the proximity of $P(t)$ to singularity. On one hand, comparing the diagonal entries on both sides of (11), we obtain the differential equation for $S(t)$

$$\frac{dS}{dt} = \operatorname{diag}(Q), \tag{16}$$

since both $Z(t)$ and $W(t)$ are skew symmetric. On the other hand, comparing the off-diagonal entries on both sides of (11), we obtain the linear system

$$q_{jk} = z_{jk}s_k + s_j w_{jk}, \tag{17}$$

$$-q_{kj} = z_{jk}s_j + s_k w_{jk}. \tag{18}$$

If $s_k^2 \neq s_j^2$, we can solve this system and obtain

$$z_{jk} = \frac{s_k q_{jk} + s_j q_{kj}}{s_k^2 - s_j^2}, \tag{19}$$

$$w_{jk} = \frac{s_j q_{jk} + s_k q_{kj}}{s_j^2 - s_k^2} \tag{20}$$

for all $j > k$. Even if $s_k^2 = s_j^2$, the existence of an ASVD guarantees that the equations must be consistent, so z_{jk} and w_{jk} still can be solved. Detailed consideration of this situation is elucidated in [22]. The basic idea is to continue differentiating (11) to obtain

$$Z^T Q + X^T \ddot{P} Y + Q W^T = \dot{Z}S + Z\dot{S} + \ddot{S} + \dot{S}W + S\dot{W}. \tag{21}$$

Picking out the terms z_{jk} and w_{kj} produces the equations

$$s_k \dot{z}_{jk} + s_j \dot{w}_{kj} + 2\dot{s}_k z_{jk} + 2\dot{s}_j w_{kj} = (X^T \ddot{P} Y)_{jk} - \sum_{i \neq k} z_{ji} q_{ik} - \sum_{i \neq j} q_{ji} w_{ki}, \tag{22}$$

$$s_j \dot{z}_{jk} + s_k \dot{w}_{kj} + 2\dot{s}_j z_{jk} + 2\dot{s}_k w_{kj} = -(X^T \ddot{P} Y)_{kj} + \sum_{i \neq j} z_{ki} q_{ij} + \sum_{i \neq k} q_{ki} w_{ji}. \tag{23}$$

As long as $\dot{s}_j \neq \dot{s}_k$, subtraction of the above two equations eliminates $\dot{z}_{jk}$ and $\dot{w}_{kj}$ and gives rise to a second relationship between z_{jk} and w_{kj}. In this way, we get a simple cross cover of the paths of the two singular values $s_j(t)$ and $s_k(t)$. If $\dot{s}_j = \dot{s}_k$ again, then it can be shown that a further differentiation of (11) will provide yet another equation to determine z_{jk} and w_{kj} as long as $\ddot{s}_j \neq \ddot{s}_k$, and the argument may continue as long as it is needed.

Once $Z(t)$ and $W(t)$ are known, the differential equations for $X(t)$ and $Y(t)$ are given, respectively, by

$$\frac{dX}{dt} = XZ, \tag{24}$$

$$\frac{dY}{dt} = YW^T. \tag{25}$$

By now we have developed a complete coordinate system $(X(t), S(t), Y(t), R(t))$ for matrices in $\mathcal{M}(\Lambda) \times \pi(R_+^n)$. The differential equations (24), (16), (25), and (9) with the relationship (10) describe how these coordinates should be varied in t to produce the steepest descent flow for the objective function F. This flow is ready to be integrated numerically by any initial value problem solvers. We have thus proposed a numerical method for solving the inverse stochastic spectrum problem.

4. Convergence. When assessing the convergence properties of the foregoing approach, we must distinguish carefully the means used to measure the convergence.

First of all, the approach fails only at two occasions—either $P(t)$ becomes singular in finite time or $F(P(t), R(t))$ converges to a nonzero constant. The former case,

detected by examining the singular values of $P(t)$, requires a restart from a new initial value with the hope of avoiding the singularity. The latter case indicates that a least squares *local* solution has been found, but that solution has not yet solved the inverse spectrum problem. A restart may help to locate an exact solution, if the prescribed spectrum is feasible, or move to another least squares approximation that may produce a different objective value.

In all cases, the function

$$G(t) := F(P(t), R(t)) \tag{26}$$

enjoys the property that

$$\frac{dG}{dt} = -\|\nabla F(P(t), R(t))\|^2 \leq 0 \tag{27}$$

along any solution curve $(P(t), R(t))$. It follows that $G(t)$ is monotone decreasing and that $\frac{dG}{dt} = 0$ only when a local stationary point of $F(P, R)$ is reached. Suppose that $P(t)$ remains nonsingular throughout the integration, an assumption that seems generic according to our experiences. Then $G(t)$ has to converge. It is in this sense that our method is globally convergent.

In [6] the coordinate matrix $P(t)$ is limited to be orthogonal; hence it is bounded and exists for all t. This constraint is not imposed on the approach discussed in the current paper. Generally there is no guarantee that $P(t)$ is bounded. However, in the case that the solution flow $(P(t), R(t))$ corresponding to a certain initial value indeed is bounded and exists for all $t \geq 0$, then we can conclude from Lyapunov's second method [4] that ω-limit points of $P(t)$ exist and that each limit point satisfies $\nabla F(P, R) = 0$. In other words, limit points of the flow are necessarily stationary points. Since the vector field always points to the steepest descent direction and other types of stationary points are unstable, any limit point reached through numerical computation will most likely be a local minimizer for F. The structure of the ω-limit set of the differential system (8) and (9) can be further analyzed in a way similar to that in [6]. For example, if the ω-limit set of a flow contains a point at which $F(P, R) = 0$, then that point is the only element in the ω-limit set. The flow hence converges to that limit point. We shall not repeat the detailed argument here. Our experiences seem to indicate that our method works reasonably well for solving the inverse spectrum problem.

5. Numerical experiment. In this section, we report some experiences of our experiment with the differential equation applied to the inverse problem. The computation is carried out by MATLAB 4.2a on an ALPHA 3000/300LX workstation. The solvers used for the initial value problem are **ode113** and **ode15s** from the MATLAB ODE SUITE [20]. The code **ode113** is a PECE implementation of Adams–Bashforth–Moulton methods for nonstiff systems. The code **ode15s** is a quasi-constant step size implementation of the Klopfenstein–Shampine family of the numerical differential formulas for stiff systems. The statistics about the cost of integration can be obtained directly from the **odeset** option built in the integrator. More details of these codes can be found in the document [20]. The reason for using these two codes is simply for convenience and illustration. Any other ODE solvers can certainly be used instead.

In our experiments, the tolerance for both absolute error and relative error is set at 10^{-12}. This criterion is used to control the accuracy in following the solution path. The high accuracy we required here has little to do with the dynamics of

the underlying vector field. We examine the output values at a time interval of 10. The integration terminates automatically when the norm of $\Delta(P,R)$ or the relative improvement of $\Delta(P,R)$ between two consecutive output points is less than 10^{-9}, indicating either a stochastic matrix with the prescribed spectrum or, in the case of an infeasible spectrum, a least squares solution has been found. So as to fit the data comfortably in the running text, we report only the case $n = 5$ and display all numbers with five digits.

Example 1. To ensure the feasibility of test data, we start with a randomly generated stochastic matrix and use its eigenvalues as the objective spectrum. To demonstrate the robustness of our approach, the initial values of the differential equations are also generated randomly. Reported below is one typical run in our experiments.

The random matrix

$$A = \begin{bmatrix} 0.0596 & 0.2586 & 0.0838 & 0.3022 & 0.2958 \\ 0.0972 & 0.2833 & 0.3559 & 0.2545 & 0.0092 \\ 0.2015 & 0.1143 & 0.3645 & 0.2669 & 0.0528 \\ 0.2637 & 0.2116 & 0.1920 & 0.0333 & 0.2994 \\ 0.1785 & 0.3138 & 0.1386 & 0.2146 & 0.1545 \end{bmatrix}$$

is stochastic. Its spectrum $\{1.0000, -0.2403, 0.1186 \pm 0.1805i, -0.1018\}$, also random but feasible, is used as the target. We note that the presence of complex-conjugate pair(s) of eigenvalues in the spectrum is quite common. Orthogonal matrices X_0, Y_0 and the diagonal matrix S_0 from the singular value decomposition $P_0 = X_0 S_0 Y_0$ of the random matrix

$$P_0 = \begin{bmatrix} 0.2002 & 0.4213 & 0.9229 & 0.7243 & 0.4548 \\ 0.6964 & 0.0752 & 0.9361 & 0.2235 & 0.0981 \\ 0.7538 & 0.3620 & 0.2157 & 0.5272 & 0.2637 \\ 0.4366 & 0.3220 & 0.8688 & 0.1729 & 0.8697 \\ 0.8897 & 0.1436 & 0.7097 & 0.5343 & 0.7837 \end{bmatrix},$$

together with the matrix $R_0 = 0.8329\mathbf{1}$, where $\mathbf{1}$ is the matrix with all entries 1, are used as the initial values for $X(t)$, $Y(t)$, $S(t)$, and $R(t)$, respectively. Figure 2 depicts the history of $F(P(t), R(t))$ throughout the integration. As is expected, $F(P(t), R(t))$ is monotone decreasing in t. The flow $P(t)$ converges to a nonnegative matrix with the prescribed spectrum that by Theorem 1.1 is converted into a stochastic matrix B:

$$B = \begin{bmatrix} 0.1679 & 0.0522 & 0.4721 & 0.0000 & 0.3078 \\ 0.1436 & 0.1779 & 0.4186 & 0.1901 & 0.0698 \\ 0.0000 & 0.1377 & 0.5291 & 0.3034 & 0.0299 \\ 0.0560 & 0.4690 & 0.2404 & 0.0038 & 0.2309 \\ 0.1931 & 0.1011 & 0.5339 & 0.1553 & 0.0165 \end{bmatrix}.$$

Note that B is not expected to be correlated to A other than the spectrum since no other information of A is used in the calculation. While the history of $F(P(t), R(t))$ is independent of the integrator used, Figure 3 indicates the number of steps taken in each interval of length 10 by the nonstiff solver **ode113** and by the stiff solver **ode15s**. Both solvers seem to work reasonably well, although the stiff solver clearly is advancing with much larger step sizes at the cost of solving implicit algebraic equations. Figure 4 summarizes the statistics of the cost when using **ode15s**. It should be pointed out that the numerical computation of the partial derivative (and

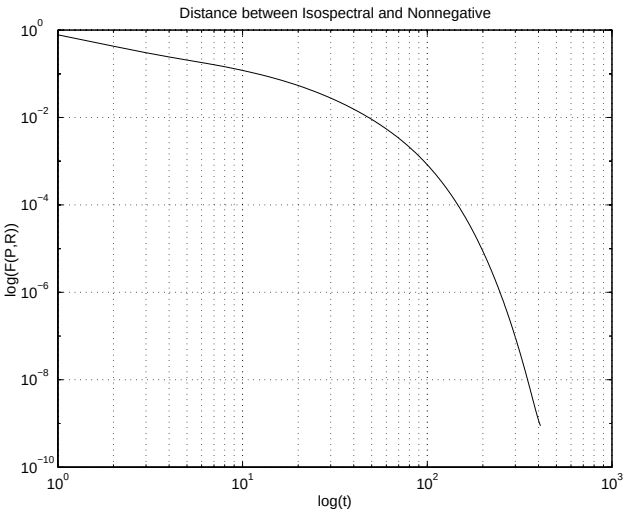


FIG. 2. *A log–log plot of* $F(P(t), R(t))$ *versus* t *for Example* 1.

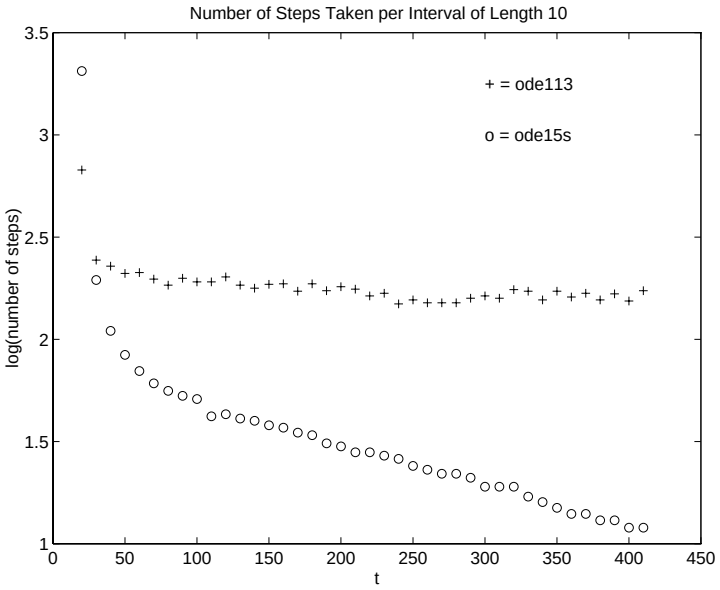


FIG. 3. *A comparison of steps taken by* **ode113** *and* **ode15s** *for Example* 1.

the related function evaluations) could have been saved if the interval of output points had been larger [20].

Suppose we merely change the initial value R_0 in the above to another random

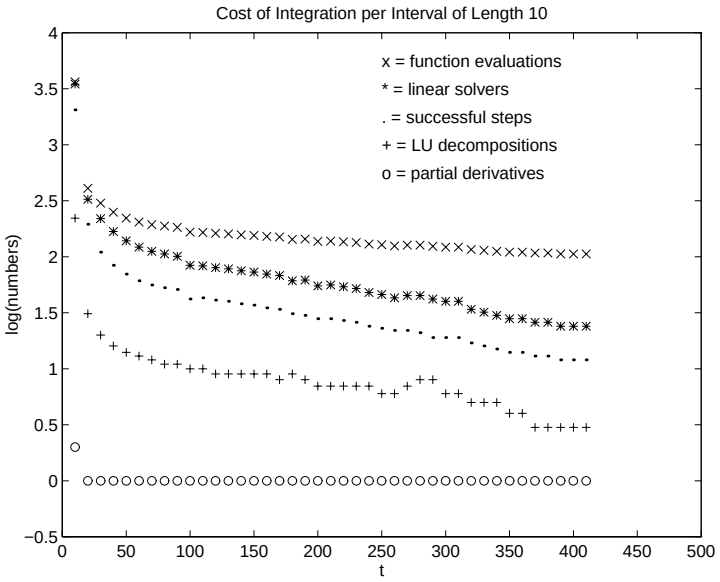


FIG. 4. *Cost of* **ode15s** *for Example* 1.

matrix:

$$R_0 = \begin{bmatrix} 0.8329 & -0.9698 & 0.2274 & 0.9466 & -0.1409 \\ -0.6222 & 0.3131 & -0.7072 & 0.6990 & -0.6490 \\ 0.5684 & 0.4914 & 0.2558 & -0.2685 & -0.0901 \\ -0.9794 & 0.6124 & -0.4724 & -0.9758 & -0.8408 \\ -0.5250 & -0.9640 & 0.0399 & -0.0852 & 0.4312 \end{bmatrix}.$$

Then the resulting stochastic matrix C becomes

$$C = \begin{bmatrix} 0.1422 & 0.0310 & 0.8267 & 0.0000 & 0.0001 \\ 0.0016 & 0.5337 & 0.2791 & 0.0756 & 0.1099 \\ 0.0000 & 0.6413 & 0.1603 & 0.0000 & 0.1984 \\ 0.2549 & 0.7019 & 0.0139 & 0.0037 & 0.0255 \\ 0.0360 & 0.6595 & 0.2178 & 0.0315 & 0.0553 \end{bmatrix},$$

illustrating the nonuniqueness of the solution for the inverse spectrum problem and also the robustness of our differential equation approach.

Example 2. In this example, we illustrate the application of our approach to the structured stochastic matrix. Suppose we want to find a stochastic matrix with eigenvalues $\{1.0000, -0.2608, 0.5046, 0.6438, -0.4483\}$. Furthermore, suppose we want the Markov chain to be such that the states form a *ring* and that each state is linked at most to its two immediate neighbors. We begin with the initial matrices

$$P_0 = \begin{bmatrix} 0.1825 & 0.7922 & 0.2567 & 0.9260 & 0.9063 \\ 0.1967 & 0.5737 & 0.7206 & 0.5153 & 0.0186 \\ 0.5281 & 0.2994 & 0.9550 & 0.6994 & 0.1383 \\ 0.7948 & 0.6379 & 0.5787 & 0.1005 & 0.9024 \\ 0.5094 & 0.8956 & 0.3954 & 0.6125 & 0.4410 \end{bmatrix}$$

and $R_0 = 0.9210\hat{\mathbf{1}}$, where

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

As we pointed out earlier, the zeros in R_0 are invariant under the integration of (8) and (9). Thus we are maintaining the ring structure while searching for the one with matched spectrum. It turns out that the stochastic matrix

$$D = \begin{bmatrix} 0.0000 & 0.3094 & 0 & 0 & 0.6906 \\ 0.0040 & 0.5063 & 0.4896 & 0 & 0 \\ 0 & 0.0000 & 0.5134 & 0.4866 & 0 \\ 0 & 0 & 0.7733 & 0.2246 & 0.0021 \\ 0.4149 & 0 & 0 & 0.3900 & 0.1951 \end{bmatrix}$$

is the limit point of the solution flow and possesses the desirable spectrum.

Example 3. By a result of Dmitriev and Dynkin [17], a complex number α with $|\arg \alpha| \leq \frac{2\pi}{n}$ is an eigenvalue of an $n \times n$ stochastic matrix if and only α lies either in the triangle $\triangle(0, 1, e^{2\pi i/n})$ or in $\triangle(0, 1, e^{-2\pi i/n})$. The result by replacing the complex-conjugate pair in the spectrum of Example 1 with another pair of complex-conjugate values in these two triangles will not alter the fact that every individual value is an eigenvalue of a certain stochastic matrix. However, whether these values are eigenvalues of the same stochastic matrix is difficult to confirm.

We experiment with, for instance, the eigenvalues $.3090 \pm 0.5000i$. Using the same initial values ($R_0 = 0.8329\mathbf{1}$) as in Example 1, we have experienced extremely slow convergence for this case. The history of $F(P, R)$ in Figure 5 clearly indicates this observation. The limit point, given by

$$E = \begin{bmatrix} 0.3818 & 0.0000 & 0.4568 & 0.0000 & 0.1614 \\ 0.5082 & 0.3314 & 0.0871 & 0.0049 & 0.0684 \\ 0.0000 & 0.0000 & 0.5288 & 0.4712 & 0.0000 \\ 0.0266 & 0.7634 & 0.0292 & 0.0310 & 0.1498 \\ 0.5416 & 0.0524 & 0.3835 & 0.0196 & 0.0029 \end{bmatrix},$$

exhibits an unexpected zero structure that we think is the cause of the slow convergence. The variation of the smallest singular value in the ASVD is plotted in Figure 6, indicating that matrices $P(t)$ stay away from singularity at a good distance. Suppose we modify the initial value to reflect the structure by simply setting the corresponding entries in the original R_0 to zero. Then the flow converges to another limit point,

$$F = \begin{bmatrix} 0.3237 & 0 & 0.4684 & 0 & 0.2079 \\ 0.4742 & 0.3184 & 0.1303 & 0.0007 & 0.0764 \\ 0 & 0.0000 & 0.5231 & 0.4769 & 0 \\ 0.0066 & 0.7536 & 0.0372 & 0.0958 & 0.1068 \\ 0.5441 & 0.0429 & 0.3959 & 0.0022 & 0.0149 \end{bmatrix},$$

at an almost equally slow pace. The spectra of both E and F agree with the specified spectrum within expected computational errors.

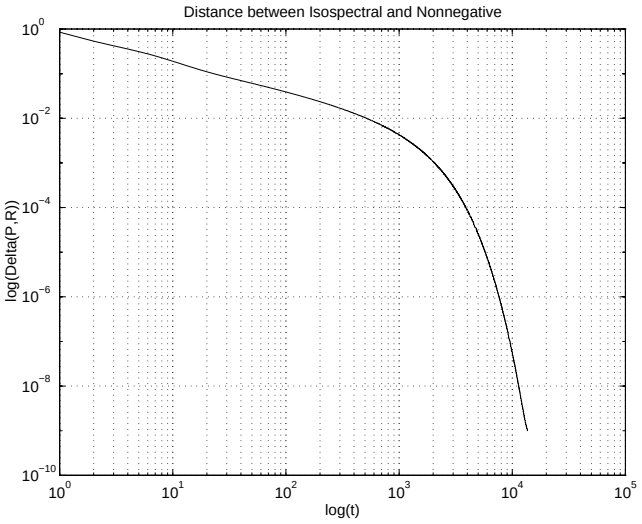


FIG. 5. *A log–log plot of* $F(P(t), R(t))$ *versus* t *for Example* 3.

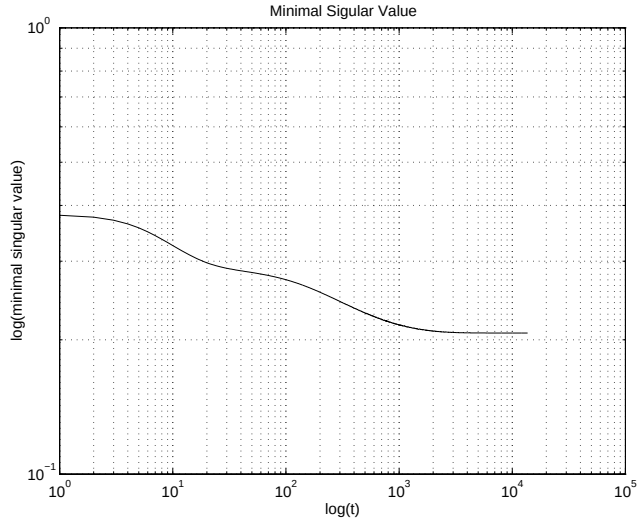


FIG. 6. *History of the smallest singular value for Example* 3.

6. Conclusion. The theory of solvability on the inverse spectrum problem for stochastic or nonnegative matrices is yet to be developed; nevertheless we have proposed an ODE approach that is capable of constructing numerically stochastic or nonnegative matrices with the desirable spectrum if the spectrum is feasible. The method is easy to implement by existing ODE solvers. The method can also be used to approximate least squares solutions or linearly structured matrices.

Acknowledgments. The authors wish to thank Bart De Moor for bringing this problem to their attention and Carl Meyer for kindly pointing them to reference [17].

REFERENCES

[1] W. W. Barrett and C. R. Johnson, *Possible spectra of totally positive matrices*, Linear Algebra Appl., 62 (1984), pp. 231–233.

[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Classics in Applied Mathematics 9, SIAM, Philadelphia, 1994.

[3] M. Boyle and D. Handelman, *The spectra of nonnegative matrices via symbolic dynamics*, Ann. of Math., 133 (1991), pp. 249–316.

[4] F. Brauer and J. A. Hohel, *Qualitative Theory of Ordinary Differential Equations*, Benjamin, New York, 1969.

[5] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols, *Numerical computation of an analytic singular value decomposition of a matrix valued function*, Numer. Math., 60 (1991), pp. 1–40.

[6] M. T. Chu and K. R. Driessel, *Constructing symmetric nonnegative matrices with prescribed eigenvalues by differential equations*, SIAM J. Math. Anal., 22 (1991), pp. 1372–1387.

[7] M. T. Chu, *Inverse eigenvalue problems*, SIAM Rev., 40 (1998), pp. 1–39.

[8] B. De Moor, *private communication*, 1994.

[9] M. Fiedler, *Eigenvalues of nonnegative symmetric matrices*, Linear Algebra Appl., 9 (1974), pp. 119–142.

[10] S. Friedland, *On an inverse problem for nonnegative and eventually nonnegative matrices*, Israel J. Math., 29 (1978), pp. 43–60.

[11] S. Friedland and A. A. Melkman, *On the eigenvalues of nonnegative Jacobi matrices*, Linear Algebra Appl., 25 (1979), pp. 239–254.

[12] G. M. L. Gladwell, *Inverse Problems in Vibration*, Martinus Nijhoff Publishers, Dordrecht, the Netherlands, 1986.

[13] D. Hershkowitz, *Existence of matrices with prescribed eigenvalues and entries*, Linear and Multilinear Algebra, 14 (1983), pp. 315–342.

[14] R. Loewy and D. London, *A note on an inverse problem for nonnegative matrices*, Linear and Multilinear Algebra, 6 (1978), pp. 83–90.

[15] F. I. Karpelevič, *On the characteristic roots of matrices with nonnegative elements*, Izv. Akad. Nauk SSSR Ser. Mat., 15 (1951), pp. 361–383 (in Russian).

[16] V. Mehrmann and W. Rath, *Numerical methods for the computation of analytic singular value decompositions*, Electron. Trans. Numer. Anal., 1 (1993), pp. 72–88.

[17] H. Minc, *Nonnegative matrices*, Wiley, New York, 1988.

[18] G. N. Oliveira, *Nonnegative matrices with prescribed spectrum*, Linear Algebra Appl., 54 (1983), pp. 117–121.

[19] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1993.

[20] L. F. Shampine and M. W. Reichelt, *The MATLAB ODE suite*, SIAM J. Sci. Comput., 18 (1997), pp. 1–22.

[21] G. W. Soules, *Constructing symmetric nonnegative matrices*, Linear Algebra Appl., 13 (1983), pp. 241–251.

[22] K. Wright, *Differential equations for the analytic singular value decomposition of a matrix*, Numer. Math., 63 (1992), pp. 283–295.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 1040–1044, October 1998

013

ON THE SINGULARITY OF LCM MATRICES*

BO-YING WANG†

Abstract. Let $S = \{x_1, x_2, \ldots, x_n\}$ be a set of distinct positive integers. The set S is called gcd-closed if it contains the greatest common divisor (x_i, x_j) of x_i and x_j for $1 \le i, j \le n$. The matrix $[S]$ is called the least common multiple (LCM) matrix on S if its i, j entry is the least common multiple $[x_i, x_j]$ of x_i and x_j. Bourque and Ligh conjectured that the LCM matrix on a gcd-closed set is invertible [*Linear Algebra Appl.*, 174 (1992), pp. 65–74]. The aim of this note is to show that this conjecture holds if $n \le 7$, but it does not hold in general when $n \ge 8$.

Key words. least common multiple, LCM matrix, greatest common divisor, gcd-closed set, singularity

AMS subject classifications. 15A36, 15A09, 11A05

PII. S0895479896313747

1. Introduction.

Let $S = \{x_1, x_2, \ldots, x_n\}$ be a set of distinct positive integers. The matrix $[S]$ having the least common multiple of x_i and x_j as its i, j entry is called the least common multiple (LCM) matrix on S. Let $(x_1, \ldots, x_k)$ denote the greatest common divisor of $x_1, \ldots, x_k$. The set S is gcd-closed if $(x_i, x_j) \in S$ for $1 \le i, j \le n$. Note that this implies $(x_{j_1}, \ldots, x_{j_k}) \in S$. Let μ denote Möbius function, i.e., $\mu(1) = 1, \mu(n) = (-1)^k$ for $n = p_1 p_2 \cdots p_k$ where $p_1, p_2, \ldots, p_k$ are distinct prime numbers and $\mu(n) = 0$ for the other n. It is well known that setting $g(n) = \frac{1}{n}\sum_{d|n} d\mu(d)$; then $g(n) = \frac{1}{n}(1-p_1)(1-p_2)\cdots(1-p_k)$ if $p_1, \ldots, p_k$ are the prime factors of n and $\sum_{d|n} g(d) = \frac{1}{n}$ (see, e.g., [1]).

The following result is due to Bourque and Ligh [2]; it generalizes Smith's result on LCM matrices [5].

LEMMA 1.1 (see [2]). *Let $S = \{x_1, x_2, \ldots, x_n\}$ be a set of distinct positive integers. If S is gcd-closed, then*

$$\det[S] = \prod_{i=1}^{n} x_i^2 \alpha_i, \quad \textit{where } \alpha_i = \sum_{\substack{d|x_i \\ d\nmid x_l \\ x_l < x_i}} g(d). \tag{1}$$

One of our results (see Theorem 1.4) is an alternative formula for calculating α_i more rapidly. We see that if S is factor-closed (i.e., S contains every divisor of any element of S); then $\det[S] = \prod_{i=1}^{n} x_i^2 g(x_i) \ne 0$, so $[S]$ is invertible (see Smith [5]). The authors of [2] conjectured that invertibility of $[S]$ holds even under the weaker property that S be only gcd-closed. This will be disproved in Example 2, while we show in Theorem 1.6 that the conjecture holds if $|S| \le 7$.

We first introduce a symmetric function $w = w(y_1, \ldots, y_k)$ on the positive integers

* Received by the editors December 6, 1996; accepted for publication (in revised form) by G.P. Styan July 7, 1997; published electronically July 17, 1998. This research was supported by the National Science Foundation of China.

http://www.siam.org/journals/SIMAX/19-4/31374.html

† Department of Mathematics, Beijing Normal University, Beijing 100875, China (bywang@sun.ihep.ac.cn).

by

$$w(\emptyset)=0,\quad w(y_1)=\tfrac{1}{y_1},\quad w(y_1,y_2)=\tfrac{1}{y_1}+\tfrac{1}{y_2}-\tfrac{1}{(y_1,y_2)},\ldots,$$
$$w(y_1,y_2,\ldots,y_k)=\sum_{t=1}^{k}(-1)^{t-1}\sum_{1\le i_1<\cdots<i_t\le k}\frac{1}{(y_{i_1},\ldots,y_{i_t})}.$$

LEMMA 1.2. *Let $y_1,y_2,\ldots,y_k$ be positive integers; then*

$$\begin{aligned}w(y_1,y_2,\ldots,y_k)=\ & w(y_1,y_2,\ldots,y_{k-1})+w(y_k)\\ & -w((y_1,y_k),(y_2,y_k),\ldots,(y_{k-1},y_k)).\end{aligned}$$

Proof. Note that $(y_1,y_2,\ldots,y_k)=((y_1,y_k),(y_2,y_k),\ldots,(y_{k-1},y_k))$ and

$$\begin{aligned}\sum_{1\le i_1<\cdots<i_t\le k}\frac{1}{(y_{i_1},\ldots,y_{i_t})}\ &=\sum_{1\le i_1<\cdots<i_t\le k-1}\frac{1}{(y_{i_1},\ldots,y_{i_t})}\\ &\quad+\sum_{1\le j_1<\cdots<j_{t-1}\le k-1}\frac{1}{(y_{j_1},\ldots,y_{j_{t-1}},y_k)}.\end{aligned}$$

We have

$$\begin{aligned}w(y_1,y_2,\ldots,y_k)&=\sum_{t=1}^{k-1}(-1)^{t-1}\sum_{1\le i_1<\cdots<i_t\le k-1}\frac{1}{(y_{i_1},\ldots,y_{i_t})}\\ &\quad+\tfrac{1}{y_k}+\sum_{t=2}^{k}(-1)^{t-1}\sum_{1\le j_1<\cdots<j_{t-1}\le k-1}\frac{1}{((y_{j_1},y_k),\ldots,(y_{j_{t-1}},y_k))}\\ &=w(y_1,y_2,\ldots,y_{k-1})+w(y_k)-w((y_1,y_k),(y_2,y_k),\ldots,(y_{k-1},y_k)).\end{aligned}$$

□

LEMMA 1.3. *Let $y_1,y_2,\ldots,y_m(m\ge 2)$ be positive integers. If $y_m|y_i$ for some $i\ne m$, then $w(y_1,y_2,\ldots,y_m)=w(y_1,y_2,\ldots,y_{m-1})$. Particularly, $w(y_1,\ldots,y_1)=w(y_1)$, $w(y_1,\ldots,y_k,1)=w(y_1,\ldots,y_k)$.*

Proof. Use induction on m.

For $m=2$, from $y_2|y_1$ we have $(y_1,y_2)=y_2$ and $w(y_1,y_2)=\frac{1}{y_1}+\frac{1}{y_2}-\frac{1}{(y_1,y_2)}=w(y_1)$.

Observe that $(y_i,y_m)=y_m$ and $(y_j,y_m)|y_m$. Using inductive hypothesis and Lemma 1.2, we obtain

$$\begin{aligned}w(y_1,y_2,\ldots,y_m)\ &=w(y_1,y_2,\ldots,y_{m-1})+w(y_m)\\ &\quad-w((y_1,y_m),\ldots,(y_{i-1},y_m),y_m,(y_{i+1},y_m),\ldots,(y_{m-1},y_m))\\ &=w(y_1,y_2,\ldots,y_{m-1})+w(y_m)-w(y_m)\\ &=w(y_1,y_2,\ldots,y_{m-1}).\end{aligned}$$

□

Now we show the following result.

THEOREM 1.4. *Let $S=\{x_1,x_2,\ldots,x_n\}$ be a set of positive integers and $x_1>x_2>\cdots>x_n$. If S is gcd-closed, then*

$$\det[S]=\prod_{i=1}^{n}x_i^2\beta_i,$$

where $\beta_i=w(x_i)-w((x_i,x_{i+1}),(x_i,x_{i+2}),\ldots,(x_i,x_n)),i=1,\ldots,n.$

Proof. By (1), we only need to prove $\alpha_i = \beta_i,\ i = 1, \ldots, n$.

For a fixed i, let $\{(x_i, x_{i+1}), (x_i, x_{i+2}), \ldots, (x_i, x_n)\} = \{y_1, \ldots, y_m\}$; then $y_j | x_i,\ j = 1, \ldots, m$, and

$$\alpha_i = \sum_{\substack{d|x_i \\ d\nmid x_l \\ l>i}} g(d) = \sum_{\substack{d|x_i \\ d\nmid (x_i, x_l) \\ l>i}} g(d) = \sum_{\substack{d|x_i \\ d\nmid y_j \\ 1\le j\le m}} g(d). \tag{2}$$

On the other hand, using $\frac{1}{k} = \sum_{d|k} g(d)$, we have

$$\begin{aligned} \beta_i &= w(x_i) - w(y_1, \ldots, y_m) \\ &= \frac{1}{x_i} - \sum_{t=1}^{m} (-1)^{t-1} \sum_{1\le i_1<\cdots<i_t\le m} \frac{1}{(y_{i_1}, \ldots, y_{i_t})} \\ &= \sum_{d|x_i} g(d) + \sum_{t=1}^{m} (-1)^t \sum_{1\le i_1<\cdots<i_t\le m} \sum_{d|(y_{i_1}, \ldots, y_{i_t})} g(d). \end{aligned} \tag{3}$$

Fix a d. If $d \nmid x_i$ then $g(d)$ occurs in neither (2) nor (3). Now assume $d|x_i$. Let $J = \{j : d|y_j\}$. If $J = \emptyset$, then the coefficients of $g(d)$ in (2) and (3) are both 1. If $|J| \ge 1$, this coefficient is in (3),

$$1 + \sum_{t=1}^{m} (-1)^t \cdot \#\{t\text{−subsets of } J\} = \sum_{t=0}^{m} (-1)^t \binom{|J|}{t} = (1-1)^{|J|} = 0,$$

as it is in (2). In this reasoning d was arbitrary so, we have shown that $\alpha_i = \beta_i$. ∎

Example 1. *Let* $S_1 = \{42, 15, 3, 1\}$; *then* S_1 *is gcd-closed, and*

$$\begin{aligned} \det[S_1] &= 42^2(w(42) - w(3)) \cdot 15^2(w(15) - w(3)) \cdot 3^2(w(3) - w(1)) \\ &= 42^2(\tfrac{1}{42} - \tfrac{1}{3}) \cdot 15^2(\tfrac{1}{15} - \tfrac{1}{3}) \cdot 3^2(\tfrac{1}{3} - 1) \\ &= 42(-13) \cdot 15(-4) \cdot 3(-2) \\ &= -196560. \end{aligned}$$

If using (1), that is,

$$\begin{aligned} \det[S_1] &= 42^2(g(42) + g(21) + g(14) + g(7) + g(6) + g(2)) \\ &\quad \cdot 15^2(g(15) + g(5)) \cdot 3^2 g(3) \\ &= 42^2(-\tfrac{2}{7} + \tfrac{4}{7} + \tfrac{3}{7} - \tfrac{6}{7} + \tfrac{1}{3} - \tfrac{1}{2}) \cdot 15^2(\tfrac{8}{15} - \tfrac{4}{5}) \cdot 3^2(-\tfrac{2}{3}) \\ &= 42(-13) \cdot 15(-4) \cdot 3(-2) \\ &= -196560. \end{aligned}$$

Example 2. $S_2 = \{30450, 174, 75, 70, 5, 3, 2, 1\}$.

Note that $30450 = 2 \cdot 3 \cdot 5 \cdot 5 \cdot 7 \cdot 29,\ 174 = 2 \cdot 3 \cdot 29,\ 75 = 3 \cdot 5 \cdot 5,\ 70 = 2 \cdot 5 \cdot 7$. We can see that S_2 is gcd-closed, but

$$\begin{aligned} \beta_1 &= \frac{1}{30450} - w(174, 75, 70) \\ &= \frac{1}{30450} - \frac{1}{174} - \frac{1}{75} - \frac{1}{70} + \frac{1}{5} + \frac{1}{3} + \frac{1}{2} - 1 \\ &= \frac{1}{30450}(1 - 175 - 406 - 435 + 6090 + 10150 + 15225 - 30450) \\ &= 0. \end{aligned}$$

Therefore $\det[S_2] = 0$.

Using (1) leads to extensive computation.

Example 2 shows that the conjecture of [2] does not hold in general. It is easy to see that, more generally, if $n \geq 8$ and we are given distinct prime numbers with $x_1 > \cdots > x_{n-8} > 30450$, then $S = \{x_1, \ldots, x_{n-8}\} \cup S_2$ is gcd-closed and $[S]$ is singular.

However we will prove that the conjecture of [2] holds if $n \leq 7$.

LEMMA 1.5. *If $y_1 \nmid y_2$, $y_2 \nmid y_1$, then $w(y_1, y_2) < 0$.*

Proof. Let $(y_1, y_2) = d, y_1 = dz_1, y_2 = dz_2$; then $z_1, z_2 \geq 2, (z_1, z_2) = 1$, and $w(y_1, y_2) = \frac{1}{d} w(z_1, z_2) = \frac{1}{d}(\frac{1}{z_1} + \frac{1}{z_2} - 1) \leq \frac{1}{d}(\frac{1}{2} + \frac{1}{3} - 1) < 0$. □

THEOREM 1.6. *Let $S = \{x_1, x_2, \ldots, x_n\}$ be a set of distinct positive integers. If S is gcd-closed and $n \leq 7$, then $[S]$ is invertible.*

Proof. Without loss of generality, we assume that $x_1 > x_2 > \cdots > x_n (n \geq 2)$ and $x_n = 1$, since S is gcd-closed.

From Theorem 1.4, we have $\beta_1 = \frac{1}{x_1} - w((x_1, x_2), (x_1, x_3), \ldots, (x_1, x_{n-1}), 1)$. Set $S' = \{(x_1, x_2), (x_1, x_3), \ldots, (x_1, x_{n-1})\}$; then $S' \subseteq S$ and $|S'| \leq 5$. We will show $\beta_1 \neq 0$.

If $S' = \phi$ or $S' = \{1\}$, then $\beta_1 = \frac{1}{x_1} - 1 < 0$. Otherwise, using Lemma 1.3 we can choose $\{y_1, \ldots, y_t\} \subseteq S'$ such that

$$\beta_1 = \frac{1}{x_1} - w(y_1, \ldots, y_t), \tag{4}$$

where $2 \leq y_i < x_1,\ y_i \nmid y_j$ for $i \neq j,\ 1 \leq i, j \leq t,\ 1 \leq t \leq 5$.

According to Lemma 1.5, $w(y_i, y_j) < 0$ for any $i \neq j$.

Note by gcd-closedness of S that if $y_i, y_j \in S', (y_i, y_j) \geq 2$, then $(y_i, y_j) \in S'$.

Now when $t = 1$, $\beta_1 = \frac{1}{x_1} - \frac{1}{y_1} < 0$. When $t = 2$, $\beta_1 = \frac{1}{x_1} - w(y_1, y_2) > 0$.

When $t \geq 3$, let us first to prove $w(y_1, y_2, y_3) < 0$. Set $(y_1, y_2, y_3) = d$; there are two cases.

Case 1. $(y_1, y_2) = d$ (or $(y_1, y_3) = d$, or $(y_2, y_3) = d$).

Then $w(y_1, y_2, y_3) = w(y_2, y_3) + \frac{1}{y_1} - \frac{1}{(y_1, y_3)} < 0$.

Case 2. $(y_1, y_2) = d_1 > d,\ (y_1, y_3) = d_2 > d,\ (y_2, y_3) = d_3 > d$.

If $d_1 = d_2$ (or $d_1 = d_3$; or $d_2 = d_3$), then $d_1 | d$, which is a contradiction. So, d_1, d_2, d_3 are distinct positive integers of S', but $d_i \notin \{y_1, y_2, y_3\},\ i = 1, 2, 3$ by (4), and this contradicts $|S'| \leq 5$. It follows that Case 2 cannot happen and $w(y_1, y_2, y_3) < 0$.

Hence when $t = 3$, we obtain $\beta_1 = \frac{1}{x_1} - w(y_1, y_2, y_3) > 0$.

For $t = 4$, let $(y_1, y_2, y_3, y_4) = d$. There are also two cases.

Case 1. There exists some y_j, e.g., y_4 such that $(y_4, y_i) = d,\ i = 1, 2, 3$.

Then $w(y_1, y_2, y_3, y_4) = w(y_1, y_2, y_3) + \frac{1}{y_4} - \frac{1}{(y_4, y_1)} < 0$, and $\beta_1 > 0$.

Case 2. For $j = 1, 2, 3, 4$ there are i_j such that $(y_j, y_{i_j}) = d_j > d$. This case is impossible, since $d_j \in S'$, $d_j \neq y_i$ (by (4)), and $|S'| \leq 5$ imply that $d_1 = d_2 = d_3 = d_4$ and $d_j | d$.

Finally, when $t = 5$, then $S' = \{y_1, \ldots, y_5\}$. So $(y_i, y_j) = 1$, for otherwise there exists a $k \neq i$ such that $(y_i, y_j) = y_k | y_i$, which contradicts (4). We obtain $w(y_1, \ldots, y_5) = \frac{1}{y_1} + \cdots + \frac{1}{y_5} - 4 < 0$, and hence $\beta_1 > 0$.

Thus, we have proved that $\beta_1 \neq 0$. Similarly, also $\beta_i \neq 0,\ i = 2, \ldots, n$. Therefore $[S]$ is invertible by Theorem 1.4. □

LEMMA 1.7. *Let $y_1, y_2, \ldots, y_m$ be positive integers and let $y_1 \geq y_j$, $(y_t, y_j) = 1$, $1 \leq t < j \leq m$. Then*

$$w(y_1, y_2, \ldots, y_m) = \begin{cases} w(y_1) & \text{if } y_j | y_1,\ j = 2, \ldots, m; \\ < 0 & \text{otherwise.} \end{cases}$$

Proof. By Lemma 1.3 we can assume that $y_1, y_2, \ldots, y_m$ are distinct.

The first case is obvious. For the others, if $m = 2$, then $w(y_1, y_2) < 0$ by Lemma 1.5; if $m \geq 3$, then $w(y_1, y_2, \ldots, y_m) = w(y_2, \ldots, y_m) + \frac{1}{y_1} - 1 = \cdots = \frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_m} - m + 1 \leq 1 + \frac{1}{2} + \frac{m-2}{3} - m + 1 = \frac{-4m+11}{6} < 0$. □

COROLLARY. *Let $S = \{x_1, x_2, \ldots, x_n\}$ be a gcd-closed set with $x_1 > x_2 > \cdots > x_n$. If $(x_t, x_j) = 1$, $2 \leq t < j \leq n$, then $[S]$ is invertible.*

Proof. By Theorem 1.4, $\beta_i = w(x_i) - w((x_i, x_{i+1}), (x_i, x_{i+2}), \ldots, (x_i, x_n))$, $i = 1, \ldots, n$. Noting that $((x_i, x_t), (x_i, x_j)) = 1$, $i+1 \leq t < j \leq n$, we obtain $\beta_i \neq 0$, $i = 1, \ldots, n$ by Lemma 1.7. It follows that $[S]$ is invertible. □

Remark. Recently, we saw that Haukkanen, Wang, and Sillanpää showed in [3] that the conjecture of [2] does not hold, in general, and a counterexample is given in the case where $n = 9$ with $x_1 = 180$. We also saw that in [4] Hong gave a counterexample in the case where $n = 8$ with $x_1 = 227700$.

REFERENCES

[1] M. AIGNER, *Combinatorial Theory*, Springer-Verlag, New York, 1979.

[2] K. BOURQUE AND S. LIGH, *On GCD and LCM matrices*, Linear Algebra Appl., 174 (1992), pp. 65–74.

[3] P. HAUKKANEN, LUN WANG, AND J. SILLANPÄÄ, *On Smith's determinant*, Linear Algebra Appl., 258 (1997), pp. 251–269.

[4] S. F. HONG, *On Bourque-Ligh conjecture of LCM matrices*, Adv. Math., 25 (1996), pp. 566–568.

[5] H. J. S. SMITH, *On the value of a certain arithmetical determinant*, Proc. London Math. Soc., 7 (1875–1876), pp. 208–212.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 1045–1073, October 1998

014

A TRUNCATED RQ ITERATION FOR LARGE SCALE EIGENVALUE CALCULATIONS*

D. C. SORENSEN† AND C. YANG†

Abstract. We introduce a new Krylov subspace iteration for large scale eigenvalue problems that is able to accelerate the convergence through an inexact (iterative) solution to a shift-invert equation. The method also takes full advantage of exact solutions when they can be obtained with sparse direct method. We call this new iteration the truncated RQ (TRQ) iteration. It is based upon a recursion that develops in the leading k columns of the implicitly shifted RQ iteration for dense matrices. Inverse-iteration-like convergence to a partial Schur decomposition occurs in the leading k columns of the updated basis vectors and Hessenberg matrices. The TRQ iteration is competitive with the rational Krylov method of Ruhe when the shift-invert equations can be solved directly and with the Jacobi–Davidson method of Sleijpen and Van der Vorst when these equations are solved inexactly with a preconditioned iterative method. The TRQ iteration is related to both of these but is derived directly from the RQ iteration and thus inherits the convergence properties of that method. Existing RQ deflation strategies may be employed directly in the TRQ iteration.

Key words. Krylov methods, Arnoldi method, Lanczos method, eigenvalues, deflation, preconditioning, restarting

AMS subject classifications. Primary, 65F15; Secondary, 65G05

PII. S0895479896305398

1. Introduction.

Recently, there have been a number of research developments in the numerical solution of large scale eigenvalue problems [21], [11], [17], [6], [19], [16], [13], [1], [5]. The state of the art has advanced considerably, and general purpose numerical software is emerging for the nonsymmetric problem [8], [4], [12], [3], [18], [10]. The development of this new general purpose software for the nonsymmetric problem is a welcomed advance. However, the methods in these packages are not able to effectively utilize a preconditioned iterative solver to implement a shift and invert spectral transformation to accelerate convergence. They all require highly accurate solutions to the shift-invert equations, and the cost of producing such accuracy with an iterative method is generally prohibitive. In this paper, we introduce a new iteration for large scale problems that is in the same spirit as the implicitly restarted Arnoldi method used in ARPACK [21], [12]. However, this new method is very amenable to acceleration of convergence with inexact (iterative) solutions to the shift-invert equations. Moreover, the algorithm introduced here can take full advantage of exact solutions when they can be obtained with a sparse direct method.

We call this new iteration the truncated RQ (TRQ) iteration. It is based upon a recursion that develops in the leading k columns of the implicitly shifted RQ iteration for dense matrices. This iteration is analogous to the well-known QR iteration, but it implicitly factors the shifted Hessenberg matrix into an RQ factorization (triangular times orthogonal) and then multiplies the factors in reverse order rather than using a QR factorization for this iteration. The main advantage in the large scale setting

*Received by the editors June 19, 1996; accepted for publication (in revised form) by D. Calvetti November 11, 1997; published electronically July 17, 1998. This work was supported in part by NSF cooperative agreement CCR-9120008 and by ARPA contract DAAL03-91-C-0047 (administered by the U.S. Army Research Office).

http://www.siam.org/journals/simax/19-4/30539.html

†Department of Computational and Applied Mathematics, Rice University, Houston, TX 77005-1892 (sorensen@caam.rice.edu, chao@caam.rice.edu).

is that inverse-iteration-like convergence occurs in the leading column of the updated basis matrix. Thus, eigenvalues rapidly converge in the leading principal submatrix of the iterated Hessenberg matrix. A partial Schur form rapidly emerges in the leading portion of the factorization. The leading principal submatrix of the iterated Hessenberg matrix becomes upper triangular with the desired eigenvalues appearing as diagonal elements.

A k-step TRQ iteration is derived by developing a set of equations that define the $k+1$st column of the updated set of basis vectors and the updated projected Hessenberg matrix that would occur if a full RQ iteration were carried out. The resulting equations have a great deal in common with the update equation that defines the rational Krylov method of Ruhe [16] and also with the projected correction equation that defines the Jacobi–Davidson method of Sleijpen and Van der Vorst [19]. The TRQ iteration is comparable to and quite competitive with the rational Krylov method when it is possible to factor and solve the shift-invert equations directly. With restarting, it is possible to define an inexact TRQ iteration that compares very favorably with the Jacobi–Davidson method. The TRQ iterations developed here are derived directly from the RQ iteration and may take advantage of all that is known about deflation strategies in the dense case. Moreover, the convergence behavior follows directly from the convergence properties of the RQ iteration.

In section 2, we derive the TRQ equations that will define the TRQ iteration and investigate the existence and uniqueness of the solution to these equations. We also introduce the formal specification of the TRQ iteration. In section 3 we turn to some implementation issues that arise when a sparse direct solution to the shift-invert equation is possible. We show that the Arnoldi relation existing in the leading k columns may be used to greatly reduce the amount of computation required to solve the TRQ equations. In section 3 we also discuss the selection of shifts to be used in the TRQ iteration when factorizations are only allowed intermittently. Also, deflation schemes are introduced. In section 4 we give several numerical examples to illustrate the convergence behavior of the TRQ iteration. We demonstrate that the convergence is cubic on symmetric problems and quadratic on nonsymmetric problems when a factorization is done at each step. We also show that the more practical alternative of factoring intermittently is quite competitive with the rational Krylov method employing the same type of shift strategy. A comparison is made with implicitly restarted Arnoldi (IRA) in the case that only one factorization is allowed, and we observe that IRA is more efficient than TRQ in this case.

In section 5, we develop the inexact TRQ iteration with restarting. Restarting is required to maintain an Arnoldi factorization and hence a Krylov relationship amongst the columns of the k-step factorization. As convergence takes place, standard deflation techniques are employed to lock converged Schur vectors, and orthogonalization against these converged vectors takes place naturally through the Arnoldi process. In some sense, this process is closely related to inverse iteration with Wielandt deflation [23, p. 596], [17, p. 117]. We illustrate an apparent numerical advantage of placing the inverse iteration within the context of the TRQ iteration and show some explicit comparisons with deflated inverse iteration indicating clear superiority of the TRQ scheme. Of course, the purpose of introducing possibly inexact solutions to the shift-invert equations is to provide for the use of preconditioned iterative solution techniques on these equations. We show numerical experiments indicating very favorable comparison with the Jacobi–Davidson method using the same iterative method for solving the update equations in both schemes. Moreover, we give some preliminary evidence that a shifted form of a standard preconditioner for the original matrix

is a satisfactory preconditioner for the update equations. Constructing a modified preconditioner for the projected update equations (as required in Jacobi–Davidson) does not seem to be necessary with inexact TRQ.

Throughout this paper, capital and lower case Latin letters denote matrices and vectors, respectively, while lower case Greek letters denote scalars. The jth canonical basis vector is denoted by e_j. The Euclidean norm is used exclusively and is denoted by $\|\cdot\|$. The transpose of a matrix A is denoted by A^T and conjugate transpose by A^H. Upper Hessenberg matrices will appear frequently and are usually denoted by the letter H. The subdiagonal elements of such Hessenberg matrices play a special role in our algorithms. The jth subdiagonal element (i.e., the $(j+1,j)$st element) of an upper Hessenberg matrix H will be denoted by β_j. The conjugate of a complex number α is denoted by $\bar{\alpha}$.

2. Truncating the RQ iteration. The implicitly shifted QR iteration is generally the method of choice for the computation of all the eigenvalues and eigenvectors of a square matrix A. Practical implementation of the algorithm begins with a complete reduction of A to upper Hessenberg form

$$AV = VH$$

with $V^HV = I$ and H upper Hessenberg. The QR iteration is then applied to H to produce a sequence of orthogonal similarity transformations

$$H^{(j+1)} \leftarrow (Q^{(j)})^H H^{(j)} Q^{(j)}, \quad V^{(j+1)} \leftarrow V^{(j)} Q^{(j)}$$

with $H^{(1)} \equiv H$, $V^{(1)} \equiv V$, and $Q^{(j)}$ implicitly constructed and applied through a "bulge chase" process that is mathematically equivalent to obtaining $Q^{(j)}$ through the QR-factorization $Q^{(j)}R^{(j)} = H^{(j)} - \mu_j I$, $j = 1, 2, \ldots$, where $\{\mu_j\}$ is a set of shifts selected as the algorithm proceeds. We use $v_i^{(j)}$ to denote the ith column of $V^{(j)}$ and $\rho_{ii}^{(j)}$ to denote the (i,i)th entry of $R^{(j)}$. It is straightforward to show that $H^{(j)}$ remains upper Hessenberg throughout and that

$$v_1^{(j+1)}\rho_{11}^{(j)} = (A - \mu_j I)v_1^{(j)} \quad \text{and} \quad (A - \mu_j I)^H v_n^{(j+1)} = v_n^{(j)}\bar{\rho}_{nn}^{(j)}.$$

Hence, the last column is an inverse iteration sequence and the first column is a power method or polynomial iteration. The implicitly restarted Arnoldi method provides a means to truncate this QR iteration and take advantage of the shifted-power-method-like convergence properties of the leading k columns of the iterated basis $V^{(j)}$ without computing the full QR factorizations. The relations between $v_i^{(j+1)}$ and $v_i^{(j)}$ for $i = 1, 2, \ldots, k$ on successive iterations are preserved in this truncated IRA iteration as if the full QR iteration had been carried out. Appropriate shift selection will force desired eigenvalues and corresponding eigenvectors to emerge in the leading portion of the factorization as the iteration proceeds.

The following are some advantages of the IRA approach: (i) the number of basis vectors stored is predetermined and fixed so that orthogonality of the Arnoldi basis vectors may be enforced numerically, and (ii) the iteration proceeds without having to compute a matrix factorization. In many situations this iteration is successful, but it can be slow to converge or fail when the desired portion of the spectrum does not have a favorable distribution with respect to the entire spectrum of A. It would be very desirable to devise a scheme that could take advantage of the inverse iteration properties of the QR iteration instead of the power iteration properties.

Algorithm 1: Implicitly shifted RQ iteration

Input: (A, V, H) with $AV = VH$, $V^H V = I$, and H is upper Hessenberg.
Output: (V, H) such that $AV = VH, V^H V = I$ and H is upper triangular.

1. for $j = 1, 2, 3, ...$ until *convergence*,
  1.1. Select a shift $\mu \leftarrow \mu_j$;
  1.2. Factor $H - \mu I = RQ$;
  1.3. $H \leftarrow QHQ^H$; $V \leftarrow VQ^H$;
2. end;

FIG. 2.1. *Implicitly shifted RQ iteration.*

An alternative to the implicitly shifted QR iteration is the implicitly shifted RQ iteration. Again, the iteration begins with a reduction to Hessenberg form, and then the iteration demonstrated in Figure 2.1 is applied.

It is easily shown that

$$(A - \mu_j I)v_1^{(j+1)} = v_1^{(j)}\rho_{11}^{(j)}.$$

Thus, the sequence $v_1^{(j)}$ in the first column is an inverse iteration sequence, and one would expect very rapid convergence of leading columns of $V^{(j)}$ to Schur vectors of A.

In the large scale setting it is generally impossible to carry out the full iteration involving $n \times n$ orthogonal similarity transformations. It would be desirable to truncate this update procedure after k steps to maintain and update only the leading portion of the factorizations occurring in this sequence. This truncation is obtained from a set of defining equations that emerge during the partial completion of an RQ step. To derive these relations, partition $V = (V_k, \hat{V})$, where V_k denotes the leading k columns of V, and let

$$H = \begin{pmatrix} H_k & M \\ \beta_k e_1 e_k^T & \hat{H} \end{pmatrix}$$

be partitioned conformably so that

$$A(V_k, \hat{V}) = (V_k, \hat{V}) \begin{pmatrix} H_k & M \\ \beta_k e_1 e_k^T & \hat{H} \end{pmatrix}. \tag{2.1}$$

Now, for a given shift μ, partially factor $H - \mu I$ to obtain

$$H - \mu I = \begin{pmatrix} H_k - \mu I_k & \hat{M} \\ \beta_k e_1 e_k^T & \hat{R} \end{pmatrix} \begin{pmatrix} I_k & 0 \\ 0 & \hat{Q} \end{pmatrix},$$

where $\hat{H} - \mu I = \hat{R}\hat{Q}$. Then

$$(A - \mu I)(V_k, \hat{V}\hat{Q}^H) = (V_k, \hat{V}) \begin{pmatrix} H_k - \mu I_k & \hat{M} \\ \beta_k e_1 e_k^T & \hat{R} \end{pmatrix}. \tag{2.2}$$

If Givens transformations were being used, for example, then to complete the RQ factorization in (2.2), one would continue applying Givens rotations from the right

using each rotation to annihilate a subdiagonal element. However, at this point of the factorization, there is a set of equations that uniquely determines the first column v_+ of the matrix $\hat{V}\hat{Q}^H$. If these equations can be formulated and solved, then the leading portion of this iteration may be obtained using just the leading $k+1$ columns $(V_k, \hat{V}e_1)$ and the leading k columns of the Hessenberg matrix H. The remaining $n-k-1$ columns of V and of H need never be formed or factored. To formulate the defining relations, equate the leading $k+1$ columns on both sides of equation (2.2) to obtain

$$(A-\mu I)(V_k, v_+) = (V_k, v)\begin{pmatrix} H_k - \mu I_k & h \\ \beta_k e_k^T & \alpha \end{pmatrix},$$

where $v = \hat{V}e_1$, $v_+ = \hat{V}\hat{Q}^H e_1$, $h = \hat{M}e_1$, and $\alpha = e_1^T \hat{R} e_1$. From this relationship, it follows that v_+ must satisfy

$$(A-\mu I)v_+ = V_k h + v\alpha \tag{2.3}$$

with $V_k^H v_+ = 0$ and $\|v_+\| = 1$ since the columns of (V_k, v_+) must be orthonormal.

These conditions may be expressed succinctly through the *TRQ equations*

$$\begin{pmatrix} A-\mu I & V_k \\ V_k^H & 0 \end{pmatrix}\begin{pmatrix} v_+ \\ -h \end{pmatrix} = \begin{pmatrix} v\alpha \\ 0 \end{pmatrix}, \quad \|v_+\| = 1. \tag{2.4}$$

In addition to these *TRQ* equations, we note that the first k columns on both sides of (2.2) are in a k-step Arnoldi relationship

$$(A-\mu I)V_k = V_k(H_k - \mu I_k) + f_k e_k^T \tag{2.5}$$

with $f_k = v\beta_k$.

The algorithm we shall develop depends upon the determination of v_+, h, and α directly from equation (2.4) rather than from the *RQ* factorization procedure. The fact that the RQ factorization exists assures that a solution to (2.4) exists even when the bordered matrix in (2.4) is singular.

The following lemmas characterize how singularity can occur in these equations. Moreover, we prove that the solution to (2.4) is unique even when the bordered matrix is singular. In the next section we show that the singular case in (2.4) is benign and easily dealt with numerically.

LEMMA 2.1. *Assume $A-\mu I$ is nonsingular (i.e., that μ is not an eigenvalue of A) and that equations* (2.4) *and* (2.5) *hold as a result of the partial RQ factorization described by* (2.2). *Then the bordered matrix*

$$B \equiv \begin{pmatrix} A-\mu I & V_k \\ V_k^H & 0 \end{pmatrix} \tag{2.6}$$

is nonsingular if and only if $V_k^H(A-\mu I)^{-1}V_k$ is nonsingular. Moreover, if $V_k^H(A-\mu I)^{-1}V_k$ is singular and z is any nonzero vector such that $V_k^H(A-\mu I)^{-1}V_k z = 0$, then $w = -(A-\mu I)^{-1}V_k z$ is nonzero and $v_+ = \frac{w}{\|w\|}$, $h = -\frac{z}{\|w\|}$, and $\alpha = 0$ satisfy the TRQ equations.

Proof. Since the RQ factorization $\hat{R}\hat{Q} = \hat{H} - \mu I$ always exists, it follows that (2.4) must hold in any case. The assumption that $A-\mu I$ is nonsingular provides the block factorization

$$B = \begin{pmatrix} I & 0 \\ V_k^H(A-\mu I)^{-1} & I \end{pmatrix}\begin{pmatrix} A-\mu I & V_k \\ 0 & -V_k^H(A-\mu I)^{-1}V_k \end{pmatrix}. \tag{2.7}$$

Clearly, B is nonsingular if and only if $V_k^H(A-\mu I)^{-1}V_k$ is nonsingular.

To establish the second part of the lemma, we show that the equation

$$\begin{pmatrix} A-\mu I & V_k \\ 0 & V_k^H(A-\mu I)^{-1}V_k \end{pmatrix}\begin{pmatrix} w \\ z \end{pmatrix} = \begin{pmatrix} v \\ V_k^H(A-\mu I)^{-1}v \end{pmatrix}\alpha \tag{2.8}$$

has a nonzero solution $(w^H, z^H)^H$ with $\alpha = 0$ if and only if and only if $V_k^H(A-\mu I)^{-1}V_k$ is singular.

To prove this, suppose first that $\alpha = 0$ and $(w^H, z^H)^H$ is a nonzero solution to (2.8). Then $V_k^H(A-\mu I)^{-1}V_k$ must be singular because $A-\mu I$ is assumed to be nonsingular. On the other hand, if we assume $V_k^H(A-\mu I)^{-1}V_k$ is singular and z is a nonzero vector such that $V_k^H(A-\mu I)^{-1}V_k z = 0$, then putting $w = -(A-\mu I)^{-1}V_k z$ will provide a nonzero solution to (2.8) with $\alpha = 0$. Moreover, w must be nonzero since z is nonzero and $(A-\mu I)^{-1}V_k$ has linearly independent columns. Therefore, $v_+ = \frac{w}{\|w\|}$, $h = -\frac{z}{\|w\|}$, and $\alpha = 0$ will satisfy the TRQ equations. ∎

Lemma 2.1 indicates that the solution to (2.4) will be unique if and only if $V_k^H(A-\mu I)^{-1}V_k$ is either nonsingular or has a one-dimensional null space. The following lemma establishes this fact and hence the uniqueness of the solution to the TRQ equations (2.4).

LEMMA 2.2. *Assume $A-\mu I$ is nonsingular and that equations* (2.4) *and* (2.5) *hold. If $G \equiv V_k^H(A-\mu I)^{-1}V_k$ is singular, then the null space of G^H is span$\{e_k\}$.*

Proof. Let $y = V_k^H(A-\mu I)^{-1}f_k$, and define $H_\mu \equiv H_k - \mu I_k$. Then

$$\begin{aligned} GH_\mu &= V_k^H(A-\mu I)^{-1}V_k H_\mu \\ &= V_k^H(A-\mu I)^{-1}[(A-\mu I)V_k - f_k e_k^T] \\ &= I_k - y e_k^T. \end{aligned} \tag{2.9}$$

If G is singular and x is any nonzero vector such that $0 = x^H G$, then (2.9) implies

$$0 = x^H G H_\mu = x^H - (x^H y)e_k^T.$$

Since $x \neq 0$, this equation implies $x^H y \neq 0$, which in turn implies that $x/(x^H y) = e_k$. Hence, $e_k^T G = 0$ and the null space of G^H is *span*$\{e_k\}$. This concludes the proof of the lemma. ∎

Finally, the following lemma indicates that exact singularity of B rarely occurs.

LEMMA 2.3. *Assume $A-\mu I$ is nonsingular and that equations* (2.4) *and* (2.5) *hold. Then $\alpha = 0$ in* (2.4) *and $V_k^H(A-\mu I)^{-1}V_k$ is singular if and only if the shift μ is an eigenvalue of $\hat{H}$ in equation* (2.1).

Proof. It is sufficient to show $V_k^H(A-\mu I)^{-1}V_k$ is singular if and only if the shift μ is an eigenvalue of $\hat{H}$ in equation (2.1). To this end, note that $V_k^H(A-\mu I)^{-1}V_k$ is singular if and only if $V_k^H(A-\mu I)^{-1}V_k z = 0$ for some $z \neq 0$. Since $(V_k, \hat{V})$ is unitary, any such z must satisfy

$$V_k z = (A-\mu I)\hat{V}g = (A-\mu I)(V_k, \hat{V})\begin{pmatrix} 0 \\ g \end{pmatrix}$$

for some nonzero vector g (i.e., $(A-\mu I)^{-1}V_k z$ must be in the range of $\hat{V}$). This implies

$$\begin{aligned} V_k z &= (V_k, \hat{V})\begin{pmatrix} H_k - \mu I_k & M \\ \beta_k e_1 e_k^T & \hat{H} - \mu I_{n-k} \end{pmatrix}\begin{pmatrix} 0 \\ g \end{pmatrix} \\ &= V_k M g + \hat{V}(\hat{H} - \mu I_{n-k})g. \end{aligned}$$

Since $(V_k, \hat{V})$ is unitary, it follows that

$$(\hat{H} - \mu I_{n-k})g = 0, \tag{2.10}$$

and since g is nonzero, this implies the singularity of $(\hat{H} - \mu I_{n-k})$.

Now, suppose that there is a nonzero g that satisfies (2.10). Observe that $Mg \neq 0$ since this would imply $A - \mu I$ is singular. Hence, the argument just given may be reversed to produce a nonzero z such that $V_k^H (A - \mu I)^{-1} V_k z = 0$, and the lemma is proved. ∎

The TRQ equations may be used to develop a truncated k-step version of the implicitly shifted RQ iteration. If a k-step Arnoldi factorization (2.5) has been obtained, then a k-step TRQ iteration may be implemented as shown in Algorithm 2 (Figure 2.2).

Algorithm 2: Truncated RQ (TRQ) iteration

Input: (A, V_k, H_k, f_k) with $AV_k = V_k H_k + f_k e_k^T, V_k^H V_k = I$, H_k upper Hessenberg.
Output: (V_k, H_k) such that $AV_k = V_k H_k, V_k^H V_k = I$ and H_k is upper triangular.

1. Put $\beta_k = \|f_k\|$ and put $v = f_k/\beta_k$;
2. **for** $j = 1, 2, 3, \ldots$ until *convergence*,
  2.1. Select a shift $\mu \leftarrow \mu_j$;
  2.2. Solve $\begin{pmatrix} A - \mu I & V_k \\ V_k^H & 0 \end{pmatrix} \begin{pmatrix} v_+ \\ -h \end{pmatrix} = \begin{pmatrix} v\alpha \\ 0 \end{pmatrix}$ with $\|v_+\| = 1$;
  2.3. Put $\alpha = 1/\|w\|, \quad v_+ = w\alpha, \quad h = -z\alpha$;
  2.3. *RQ* Factor $\begin{pmatrix} H_k - \mu I_k & h \\ \beta_k e_k^T & \alpha \end{pmatrix} = \begin{pmatrix} R_k & r \\ 0 & \rho \end{pmatrix} \begin{pmatrix} Q_k & q \\ \sigma e_k^T & \gamma \end{pmatrix}$;
  2.4. $V_k \leftarrow V_k Q_k^H + v_+ q^H$;
  2.5. $\beta_k \leftarrow \sigma e_k^T R_k e_k$; $v \leftarrow v_k \bar{\sigma} + v_+ \bar{\gamma}$;
  2.6. $H_k \leftarrow Q_k R_k + \mu I_k$;
3. **end**;

FIG. 2.2. *The truncated RQ iteration.*

The key idea here is to determine the $k+1$st column v_+ of the updated matrix V and the $k+1$st column of H that would have been produced in the RQ iteration by solving the linear system (2.4). Then, the iteration is completed through the normal RQ iteration. As eigenvalues converge, the standard deflation rules of the RQ iteration may be applied. Orthogonality of the basis vectors is explicitly maintained through accurate solution of the defining equation. Moreover, even if the accuracy of this solution is relaxed, orthogonality may be enforced explicitly through the orthogonalization scheme developed in [7]. We shall refer to this as the DGKS procedure. Potentially, the linear solve indicated at Step 2.2 of Algorithm 2 could be provided by a straightforward block elimination scheme. However, considerable refinements to this scheme are possible due to the existing k-step Arnoldi relationship (2.5). This will be discussed in the next section.

3. Implementation issues. In this section, we address some practicalities associated with efficient implementation of the TRQ iteration.

3.1. Solving the TRQ equations. The truncated RQ iteration described in the previous section will only be effective in the large scale setting if there is an efficient means for solving the TRQ equations. Recall that A, H_k, V_k, and $f_k = v\beta_k$ are in a k-step Arnoldi relation (3.1) so that

$$(A - \mu I)V_k = V_k(H_k - \mu I_k) + f_k e_k^T. \tag{3.1}$$

Rescaling the right-hand side of the system (2.4) leads to

$$\begin{pmatrix} A - \mu I & V_k \\ V_k^H & 0 \end{pmatrix} \begin{pmatrix} w \\ z \end{pmatrix} = \begin{pmatrix} f_k \\ 0 \end{pmatrix}. \tag{3.2}$$

If we put $d = (A - \mu I)^{-1} f_k$ and $y = V_k^H d$, then block Gaussian elimination leads to solving the equations

(i) $V_k^H (A - \mu I)^{-1} V_k z = y$,
(ii) $(A - \mu I)w = f_k - V_k z$.

If $A - \mu I$ is nonsingular, these two equations together with equation (3.1) may be used to derive a solution to equation (3.2) with just a single linear solve. It is not necessary to solve a blocked system of k equations as the straightforward application of block Gaussian elimination described in the previous section would indicate. Moreover, this efficient solution scheme does not depend on determining the singularity of the TRQ equations (2.4) in any way. The underlying theory is developed with the following lemma.

LEMMA 3.1. *Assume $A - \mu I$ is nonsingular, and define $G \equiv V_k^H (A - \mu I)^{-1} V_k$ and $H_\mu \equiv (H_k - \mu I_k)$. There is a vector s such that either*

$$(I_k - H_\mu G)s \neq 0 \quad \text{or} \quad e_k^T G s \neq 0. \tag{3.3}$$

For any such s, put

$$w \equiv (I - V_k V_k^H)(A - \mu I)^{-1} V_k s.$$

Then $w \neq 0$ and a solution v_+, h, α to (2.4) is given by

$$v_+ = w/\|w\|, \quad h = (I_k - H_\mu G)s/\|w\|, \quad \alpha = -\beta_k e_k^T G s/\|w\|.$$

Proof. If $e_k^T G s = 0$ for all vectors s, then the matrix $H_\mu G$ is singular and there must be a nonzero vector s such that $(I_k - H_\mu G)s \neq 0$. Therefore, there is a k-dimensional vector s that satisfies either $\theta \equiv e_k^T G s \neq 0$ or $(I_k - H_\mu G)s \neq 0$.

For any such s, put $w \equiv (I - V_k V_k^H)(A - \mu I)^{-1} V_k s$. Observe that

$$\begin{aligned} (A - \mu I)w &= (A - \mu I)(I - V_k V_k^H)(A - \mu I)^{-1} V_k s \\ &= V_k s - (A - \mu I) V_k G s \\ &= V_k s - [V_k H_\mu + f_k e_k^T] G s \\ &= V_k (I_k - H_\mu G)s - f_k \theta. \end{aligned} \tag{3.4}$$

The conditions on s assure that the right-hand side of (3.4) is nonzero. It follows that $w \neq 0$ and that

$$(A - \mu I)v_+ = V_k h + v\alpha,$$

where $v_+ = w/\|w\|$, $h = (I_k - H_\mu G)s/\|w\|$, and $\alpha = -\beta_k \theta/\|w\|$. ∎

Remark 1. Our original motivation for developing Lemma 3.1 was to handle the case when μ is an eigenvalue of H_k. A particular choice of s for this case is to put $s = q$ where $q^H H_\mu = 0$ and $q^H q = 1$. Then

$$q^H(I_k - H_\mu G)s = q^H s = q^H q = 1.$$

The conditions of Lemma 3.1 are clearly satisfied with this choice of s. However, we do not use this choice in practice.

Remark 2. The most general form of selecting a right-hand side for constructing w is to take

$$w \equiv (I - V_k V_k^H)(A - \mu I)^{-1}(V_k t + f_k \eta),$$

where $s \equiv t - H_\mu e_k \eta$ is chosen to satisfy the conditions of Lemma 3.1. To see this, observe that

$$\begin{aligned} V_k t + f_k \eta &= V_k s + [V_k H_\mu + f_k e_k^T] e_k \eta \\ &= V_k s + (A - \mu I) V_k e_k \eta. \end{aligned}$$

Hence,

$$(I - V_k V_k^H)(A - \mu I)^{-1}(V_k t + f_k \eta) = (I - V_k V_k^H)(A - \mu I)^{-1} V_k s.$$

Thus, there is no mathematical reason to include the term $f_k \eta$, but the additional freedom may eventually have some numerical consequences that are not apparent at the moment. Note that when the shift μ is an eigenvalue of H_k then the combination of $t = 0, \eta = 1$ is prohibited because the corresponding vector s does not satisfy either of the conditions (3.3) required for constructing the solution in Lemma 3.1. The parameters t and η here are obviously related to the corresponding parameters appearing in the rational Krylov subspace (RKS) method. It is interesting to note that the choice $t = 0, \eta = 1$ is also prohibited in RKS when μ is an eigenvalue of H_k.

Remark 3. An alternative to forming h as described in Lemma 3.1 is to form w as described above and normalize to get $v_+ = w/\|w\|$. Then, construct h and α using the DGKS procedure to orthogonalize the vector $(A - \mu I)v_+$ against V_k and f_k, respectively. Thus,

$$h \leftarrow V_k^H (A - \mu I) v_+ = V_k^H A v_+, \quad \alpha \leftarrow f_k^H (A - \mu I) v_+ / \|f_k\|.$$

Lemma 3.1 justifies Algorithm 3 to solve the TRQ equations. Once again, we remark that the DGKS procedure may be used at Steps 2, 3, and 4 of Algorithm 3 to assure that both $V_k^H v_+ = 0$ and $(A - \mu I)v_+ = V_k h + v\alpha$ to working accuracy. For relatively small values of k, the main computational effort is the solution of the equation $(A - \mu I)w = V_k t + f_k \eta$. As mentioned in Remark 2, there may be advantageous choices of t and η to overcome inaccuracies due to ill-conditioning when μ is very nearly an eigenvalue of A. We used $t = e_k$ and $\eta = 0$ in all of the experiments reported in section 4. This choice seemed to perform consistently well as compared to many of the obvious choices such as taking t to be an eigenvector of H_k. Finally, it is clear that incremental rescaling may be introduced as in inverse iteration to avoid overflow and that the scalar θ appearing in the proof of Lemma 3.1 need not be computed explicitly.

Algorithm 3: Direct Solution of the TRQ Equations

Input: (A, V_k, H_k, f_k, μ) with $AV_k = V_k H_k + f_k e_k^T$, $V_k^H V_k = I$ and $V_k^H f_k = 0$.
Output: (v_+, h, α) such that $(A - \mu I)v_+ = V_k h + f_k \alpha$, $V_k^H v_+ = 0$ and $\|v_+\| = 1$.

1. Choose t and η and solve $(A - \mu I)w = V_k t + f_k \eta$;
2. $y \leftarrow V_k^H w$;
3. $w \leftarrow w - V_k y$;
4. $v_+ \leftarrow \frac{w}{\|w\|}$; $\alpha \leftarrow f_k^H (A - \mu I)v_+ / \|f_k\|$; $h \leftarrow V_k^H A v_+$;

FIG. 3.1. *Direct solution of the TRQ equations.*

The formulation just developed is appropriate when a sparse direct factorization of $A - \mu I$ is feasible. When this is not the case we must resort to an iterative scheme. For an iterative scheme, there may be an advantage to solving the projected equation

$$(I - V_k V_k^H)(A - \mu I)(I - V_k V_k^H)\hat{w} = f_k$$

and putting

$$v_+ \leftarrow \frac{w}{\|w\|},$$

where $w = (I - V_k V_k^H)\hat{w}$. This is mathematically equivalent to solving the TRQ equations. The advantage here is that the matrix

$$(I - V_k V_k^H)(A - \mu I)(I - V_k V_k^H)$$

is most likely to be much better conditioned than $A - \mu I$ when μ is near an eigenvalue of A. A projected equation of this form plays a key role in the Jacobi–Davidson method recently developed in [19], [20], [9]. It also provides a means for allowing inaccurate solutions and preconditioning as we shall discuss later in section 5.

3.2. Selection of shifts. Another important issue to be addressed in the TRQ iteration is the selection of shifts. Various options are available. They lead to different convergence behavior. We discuss only a few simple options below. The tradeoffs and comparison to other algorithms will also be discussed in section 4.

The simplest strategy is to use a fixed shift μ throughout the TRQ iteration. This shift is referred to as the *target* shift in the following discussion. In this case, a single matrix factorization of $A - \mu I$ may be used repeatedly to get inverse power method type of convergence. However, if the ratio

$$\sigma = \frac{|\lambda_j - \mu|}{|\lambda_{j+1} - \mu|} \tag{3.5}$$

is close to 1, the approximation to λ_j converges extremely slowly. In section 5, we compare this approach with the shifted and inverted IRA. It is observed that the shifted and inverted IRA is often more efficient in obtaining a few eigenvalues near a prescribed shift.

At the other extreme, we could adjust the shift at each iteration to enhance the rate of convergence. Eigenvalues of H_k are natural candidates for the shift. They

provide the best approximations to eigenvalues of A from the subspace spanned by the columns of V_k and are referred to as the Ritz values. Before each TRQ update, we compute the Ritz values and choose the one closest to the target shift as the next shift. A converged Ritz value should not be selected as a shift.

This choice of shift usually leads to a quadratic or cubic convergence rate. However, this rapid convergence is obtained at the cost of factoring a matrix at each iteration. It is observed from our experiments that Ritz values tend to jump around during the early stage of the TRQ iteration. Thus, the target shift is used during the first few iterations until Ritz values start to settle down.

A compromise between the first and the second choice is to use a fixed shift until an eigenvalue has converged. Another possibility is to use each shift for (at most) a fixed number of iterations. In either case, the best Ritz value that has not yet converged may be selected as the next shift. Rapid convergence is generally obtained with this strategy. The cost for matrix factorization is reduced in comparison with the second approach. It will be shown in section 5 that this scheme is very competitive with the rational Krylov method of Ruhe [15], [14], [16].

Finally, the leading k-columns of the implicitly shifted RQ iteration may be obtained by selecting the same set of shifts as the full dense algorithm if desired. For example, if the elements of the matrix H are denoted by γ_{ij}, we could use γ_{11} as the shift. This corresponds to the Rayleigh quotient shift in the RQ algorithm. Another alternative is the Wilkinson shift. This is defined to be the eigenvalue of the leading 2×2 matrix

$$\begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}$$

that is the nearest to γ_{11}. These strategies may be used when no target shift is given in advance or when the TRQ iteration is used in conjunction with a deflation scheme to compute the full spectrum of A.

Once the shift is chosen, an RQ update as described in Steps 2.3 through 2.6 of Algorithm 2 is taken. Clearly, it can be done explicitly, but there may be some advantage to an implicit application. An implicit shift application is straightforward since

$$\begin{pmatrix} H_k - \mu I_k & h \\ \beta_k e_k^T & \alpha \end{pmatrix} = \begin{pmatrix} H_k & h \\ \beta_k e_k^T & \tilde{\alpha} \end{pmatrix} - \mu \begin{pmatrix} I_k & 0 \\ 0 & 1 \end{pmatrix},$$

where $\tilde{\alpha} = \alpha + \mu$. Thus, the standard bulge-chase implementation of an RQ sweep corresponding to the shift μ may be applied to the matrix

$$\begin{pmatrix} H_k & h \\ \beta_k e_k^T & \tilde{\alpha} \end{pmatrix}.$$

Finally, when the matrix A is real nonsymmetric, we would like to perform the TRQ iteration in real arithmetic. However, there seems to be no simple analog to the double shifting strategy used in the QR algorithm. Applying double shifts implicitly in the TRQ iteration is possible. However, the corresponding TRQ equation involves $\hat{A} = (A - \bar{\mu} I)(A - \mu I)$, and more work is required to solve this equation. It is still questionable whether a truncated double implicit shifting strategy should be used in practice. Therefore, we shall not present the details here. A double shift algorithm that involves solving $\hat{A}w = v$ may be found in [22].

3.3. Deflation. As discussed earlier, in each TRQ iteration the TRQ equation (2.4) is solved so that a truncated Hessenberg reduction of the form

$$A(V_k, v_+) = (V_k, v) \begin{pmatrix} H_k & h \\ \beta_k e_k^T & \alpha \end{pmatrix} \tag{3.6}$$

is maintained. As the TRQ iteration proceeds, the leading subdiagonal elements of H_k become small. Usually, they will become small in order (from top down) but occasionally this convergence happens further down the subdiagonal. When the magnitude of a subdiagonal element β_j falls below some numerical threshold, it is set to zero and the matrix H_k is split to give

$$H_k = \begin{pmatrix} H_j & M \\ 0 & \hat{H}_{k-j} \end{pmatrix}.$$

The first j columns of V_k form a basis for an invariant subspace of A, and j eigenvalues of A may be extracted from H_j. The deflation technique used in the QR algorithm can be applied here to obtain subsequent eigenvalues. We rewrite (3.6) as

$$(A-\mu I)(V_j, \hat{V}_{k-j}, v^+) = (V_j, \hat{V}_{k-j}, v) \begin{pmatrix} H_j - \mu I_j & M & h_1 \\ 0 & \hat{H}_{k-j} - \mu I_{k-j} & h_2 \\ 0 & \beta_k e_{k-j}^T & \alpha \end{pmatrix}, \tag{3.7}$$

where

$$V_k = (V_j, \hat{V}_{k-j}) \quad \text{and} \quad h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}$$

have been partitioned conformably with V_j representing the leading j columns of V_k and h_1 representing the first j components of h.

An upper triangular matrix $\hat{R}$ and an orthogonal matrix $\hat{Q}$ of the form

$$\hat{R} = \begin{pmatrix} R_2 & r \\ 0 & \rho \end{pmatrix}, \quad \hat{Q} = \begin{pmatrix} Q_2 & q \\ \sigma e_{k-j}^T & \gamma \end{pmatrix}$$

are constructed such that

$$\begin{pmatrix} \hat{H}_{k-j} - \mu I_{k-j} & h_2 \\ \beta_k e_{k-j}^T & \alpha \end{pmatrix} = \hat{R}\hat{Q}.$$

Multiplying (3.7) from the right by $\tilde{Q}^H = \begin{pmatrix} I_j & \\ & \hat{Q}^H \end{pmatrix}$ yields

$$(A-\mu I)(V_j, \hat{V}_{k-j}^+, \hat{v}_+) = (V_j, \hat{V}_{k-j}, v) \begin{pmatrix} H_j - \mu I_j & \hat{M} & \hat{h}_1 \\ 0 & R_2 & r \\ 0 & 0 & \rho \end{pmatrix},$$

where $\hat{V}_{k-j}^+ = \hat{V}_{k-j} Q_2^H + v_+ q^H$, $\hat{v}_+ = \bar{\sigma}\hat{V}_{k-j} e_{k-j} + \bar{\gamma} v_+$, $\hat{M} = M Q_2^H + h_1 q^H$, and $\hat{h}_1 = \bar{\sigma} M e_{k-j} + \bar{\gamma} h_1$. Note that the V_j and H_j are not modified during the deflation.

The next cycle of TRQ iteration starts with the selection of a new shift. The roles of $\hat{H}_{k-j}$,$\hat{V}_j$ and $\hat{v}_+$ are replaced by $\hat{H}_{k-j}^+ = Q_2 R_2 + \mu I_{k-j}$, $\hat{V}_j^+$ and $\hat{v}_+$, respectively.

If the subdiagonal elements of H_k converge to zero in order (from top to bottom,) a partial Schur form

$$AV_j = V_j R_j$$

is obtained. Of course, when a subdiagonal β_j approaches zero out of order, then the splitting described in equation (3.7) above will still yield a partial Schur form since the Schur form of $H_j Q_j = Q_j R_j$ can be used to make an explicit transformation.

4. Numerical examples. In this section, we evaluate the cost and performance of the TRQ iteration. We first show an example indicating that the convergence rate of TRQ is exactly the same as that of the RQ iteration when the TRQ equations (2.4) are solved exactly. Comparisons will be made with the shifted and inverted IRA, the RKS method, and the recently proposed Jacobi–Davidson QR (JDQR) method [9]. We show that if the shift is fixed, TRQ does not provide much advantage over the shifted and inverted IRA. However, if the shifts are allowed to change during the iteration, TRQ often performs better than IRA in terms of number of iterations and is competitive with the RKS and the JDQR algorithms. Numerical examples will be presented to demonstrate the performance of the algorithm. All numerical experiments are performed using MATLAB 4.2 on a SUN-SPARC 2.

4.1. Convergence rate of TRQ. The rate of convergence of TRQ follows from that of the full RQ iteration. For certain choices of shifts, it is cubic for symmetric eigenvalue problems and quadratic for nonsymmetric problems. In fact, if the Arnoldi iteration with the starting vector v_0 is used to produce the Hessenberg reduction required by Algorithm 1 as an input, the first k eigenvalues appearing on the diagonal of the output triangular matrix will be exactly the same as the those computed by TRQ with the same starting vector.

In the following, we present an example that verifies the fast convergence of TRQ. We choose to work with a standard 5-point discrete Laplacian defined on $[0,1] \times [0,1]$ with zero Dirichlet boundary conditions. For simplicity, the 100 by 100 symmetric matrix is scaled by h^2, where $h = 1/101$ is the mesh size of the discretization. We are interested in 4 eigenvalues with the smallest magnitude. The size of the Arnoldi factorization used in the TRQ iteration is set to be 5 ($k = 5$.) In each TRQ iteration, eigenvalues of the 5×5 tridiagonal matrix H_5 defined in Step 2.6 of Algorithm 2 are computed. The one closest to zero that has not yet converged is chosen as the next shift μ. Table 4.1 lists the subdiagonal element β_j ($j = 1, 2, 3, 4$) of H_5 at each iteration. Once $|\beta_j|/(|H_{j,j}| + |H_{j+1,j+1}|)$ drops below a prescribed tolerance of 10^{-15}, we set β_j to zero. Clearly, the first eigenvalue converges cubically and the second one shows cubic convergence rate after the first one has converged. At the end of the 12th iteration, all four eigenvalues

$$\begin{aligned}
\lambda_1 &= 0.16203, \\
\lambda_2 &= 0.39851, \\
\lambda_3 &= 0.39851, \\
\lambda_4 &= 0.63499
\end{aligned}$$

are found. The convergence criterion here was a tolerance of 10^{-15} in the test for declaring a subdiagonal element to be zero. The computed direct residuals for all converged eigenpairs were on the order of 10^{-15}. The multiplicity of the eigenvalue 0.39851 is detected.

TABLE 4.1
Convergence history of the 4 computed eigenvalues of a two-dimensional Laplacian.

Iteration	μ	β_1	β_2	β_3	β_4
1	0.18638	2.31×10^{-2}	2.18×10^{0}	1.91×10^{0}	1.58×10^{0}
2	0.16204	2.33×10^{-7}	6.23×10^{-1}	1.84×10^{0}	2.18×10^{0}
3	0.16203	1.11×10^{-21}	2.10×10^{-1}	1.36×10^{0}	1.84×10^{0}
4	0.44417	0	7.92×10^{-2}	1.27×10^{-1}	1.55×10^{0}
5	0.39857	0	1.36×10^{-5}	3.83×10^{-2}	7.24×10^{-1}
6	0.39851	0	4.08×10^{-17}	1.36×10^{-1}	9.47×10^{-2}
7	0.40410	0	0	1.34×10^{-2}	3.14×10^{-2}
8	0.39851	0	0	3.84×10^{-8}	4.24×10^{-2}
9	0.39851	0	0	8.58×10^{-21}	5.71×10^{-2}
10	0.63614	0	0	0	2.15×10^{-3}
11	0.63499	0	0	0	1.52×10^{-10}
12	0.63499	0	0	0	1.88×10^{-28}

TABLE 4.2
Comparison of computational work and storage between TRQ and IRA. We assume that k eigenvalues closest to the shift σ are of interest. An Arnoldi factorization of length k is maintained in TRQ, and $p(\geq 1)$ shifts are applied in each IRA iteration (i.e., an Arnoldi factorization of length $k+p$ is maintained). We use MATVEC to denote the matrix vector multiplication used in TRQ, and use SOLVE to indicate the cost of solving a linear system in both TRQ and IRA. The operation GEMV refers to dense matrix vector multiplications needed in carrying out Arnoldi factorization. The RQ or QR update refers to the bulge chase process used in both algorithms.

	TRQ	IRA
Initialization cost	MATVEC (k times): *variable* GEMV: $O(nk^2)$ Factorization: *variable*	SOLVE ($k+p$ times): *variable* GEMV: $O(n(k+p)^2)$ Factorization: *variable*
Cost per iteration	SOLVE: *variable* Shift selection $O(k^3)$ RQ update: $O(nk+k^2)$	SOLVE (p times): *variable* GEMV: $O(n(k+p)^2)$ Shift selection: $O((k+p)^3)$ QR update: $O(n(k+p)+(k+p)^2)$
Storage	$O(n(k+1)+(k+1)^2)$	$O(n(k+p+1)+(k+p+1)^2)$

4.2. Comparison with IRA. It is mentioned in section 3.2 that a simple way of selecting a shift in Step 2.1 of Algorithm 2 is to use a fixed shift throughout the TRQ iteration. Besides its simplicity, this strategy may also reduce the computational cost when factoring $A - \mu I$ is expensive. However, as one may expect, the convergence rate of each desired eigenvalue is typically linear in this case. When the ratio σ defined in (3.5) is close to 1, slow convergence is usually observed. In the following, we compare this variant of the TRQ algorithm with the shifted and inverted IRA since both algorithms factor the matrix $A - \mu I$ only once. It is shown in Table 4.2 that TRQ requires slightly less work and storage per iteration. However, our numerical experiments often show that the shifted and inverted IRA converges faster than TRQ with the same shift. An example is presented below to demonstrate this phenomenon. The problem involves the two-dimensional Laplacian used in the previous section. The four smallest eigenvalues are sought. We placed the target shift at zero and ran TRQ with $k = 5$ (TRQ(5)). The results are compared with IRA with $k = 4$, $p = 1$ (IRA(1)), and IRA with $k = 4$, $p = 4$ (IRA(4).) The value of p indicates the number of shifts used in the IRA iteration [21]. Since the ratio $\rho = |\lambda_1|/|\lambda_2|$ is close to 1, we expect TRQ to converge slowly. In Table 4.3, we list the converged eigenvalues and the number of linear systems solved before each eigenvalue has converged. One way to accelerate the TRQ iteration is to increase the size of the Arnoldi factorization.

TABLE 4.3
Comparison of IRA and TRQ on a two-dimensional Laplacian.

Eigenvalue	TRQ(5)	IRA(1)	IRA(4)
0.16203	36	11	6
0.39851	79	16	7
0.39851	161	26	8
0.63499	186	40	11

TABLE 4.4
Comparison of TRQ(k) with different values of k.

k	No. of linear solves
5	186
10	132
15	132

The motivation is to take advantage of large gaps that may exist in the unwanted portion of the spectrum. However, the gain is usually not significant unless such gaps are large enough. In Table 4.4, we compare the total number of linear solves used in finding the four desired eigenvalues of the two-dimensional Laplacian with different k values. We observe that as k increases, the number of linear solves required in TRQ does not always decrease. Clearly, one does not want to use a k that is too large for this will increase the computational cost.

4.3. Comparison with RKS. The convergence rate of TRQ may be improved if shifts are chosen to be the best eigenvalue approximations from the subspace spanned by columns of V_k. However, this scheme requires factoring a matrix $A - \mu_j I$ at each iteration. To reduce the overall cost of TRQ, the third shift selection strategy discussed in section 3.2 may be used; i.e., a shift is used repeatedly until either a Ritz value has converged or a fixed number of iterations has occurred. Then a new shift is selected. This strategy is also employed in the RKS introduced by Ruhe [15], [14], [16]. In this section, we show by numerical example that TRQ is competitive with RKS.

The basic recursion involved in RKS [15] may be characterized by the equation

$$AV_{k+1}\hat{H}_k = V_{k+1}\hat{G}_k,$$

where V_{k+1} is n by $k+1$, $\hat{H}_k$ and $\hat{G}_k$ are $k+1$ by k, and $V_{k+1}^H V_{k+1} = I_{k+1}$. We denote the jth column of V_{k+1}, $\hat{H}_k$, and $\hat{G}_{k+1}$ by v_j, h_j, and g_j, respectively. They are produced by a sequence of Arnoldi-like steps shown in Figure 4.1.

The choice of t_j is arbitrary, but $t_j = e_j$ is recommended. The subspace spanned by the columns of V_k do not form a Krylov subspace, and approximate eigenvalues may be obtained by solving the generalized eigenvalue problem

$$G_k s = \mu H_k s, \tag{4.1}$$

where G_k and H_k are the submatrices consisting of the first k rows of $\hat{G}_k$ and $\hat{H}_k$, respectively. The convergence of each Ritz value can be monitored by the estimate derived in [15]. Deflation must be done properly [16] to avoid missing multiple eigenvalues. The cost of RKS per iteration is listed in Table 4.5.

It is mentioned in [16] that a large basis is needed when the eigenvalue problem is ill-conditioned. Thus, reorthogonalization becomes expensive. Purging and restarting

Rational Krylov subspace (RKS) iteration

Input: (A, v_1) such that $\|v_1\| = 1$.
Output: $(V_{k+1}, \hat{H}_k, \hat{G}_k)$ such that $AV_{k+1}\hat{H}_k = V_{k+1}\hat{G}_k, V_{k+1}^H V_{k+1} = I$, and H_{k+1} is upper Hessenberg.

1. Choose $t_1 = e_1$;
2. $V_1 \leftarrow (v_1)$; $\hat{H}_0 = (\,)$; $\hat{G}_0 = (\,)$;
3. **for** $j = 1, 2, 3, ..., k$.
  3.1. Choose a shift μ_j;
  3.2. $w_{j+1} \leftarrow (A - \mu_j I)^{-1}(V_j t_j)$;
  3.3. $h_j \leftarrow V_j^H w_j$; $\hat{H}_j \leftarrow (\hat{H}_{j-1},\ h_j)$;
  3.4. $g_j \leftarrow h_j\mu_j + t_j$; $\hat{G}_j \leftarrow (\hat{G}_{j-1},\ g_j)$;
  3.5. $w_{j+1} \leftarrow w_{j+1} - V_j h_j$; $\beta_j = \|w_{j+1}\|$;
  3.6. $\hat{H}_j \leftarrow \begin{pmatrix} \hat{H}_j \\ \beta_j e_j^T \end{pmatrix}$; $\hat{G}_j \leftarrow \begin{pmatrix} \hat{G}_j \\ \mu_j\beta_j e_j^T \end{pmatrix}$;
  3.7. $v_{j+1} \leftarrow w_{j+1}/\beta_j$; $V_{j+1} \leftarrow (V_j,\ v_{j+1})$;
  3.8. Choose a vector t_{j+1};
4. **end**

FIG. 4.1. *Rational Krylov subspace iteration.*

TABLE 4.5
The cost of the RKS iteration. The value of k is usually much larger than the number of desired eigenvalues k_d. Again, SOLVE refers to solving a linear system in Step 3.2 of the algorithm. The operation GEMV refers to dense matrix vector multiplications needed in carrying out the RKS factorization. Ritz approximation refers to solving the generalized eigenvalue problem $H_k s = \mu G_k s$.

Operation	Cost
Factorization(intermittently)	*variable*
SOLVE	*variable*
GEMV	$O(nk^2)$
Ritz approximation	$O(k^3)$
Purging & restart	$O(nkk_d + k^4)$
Storage	$O(n(k+1) + 2(k+1)^2)$

have been proposed in [16]. However, these schemes are still experimental and not well understood. In contrast, the size of V_k is fixed during the TRQ iteration, and the update is done by an orthogonal transformation. The convergence can be monitored by checking the magnitude of subdiagonal elements of H_k. Deflation is built into the TRQ iteration, and eigenvalues with multiplicity greater than one cause no difficulty. At convergence, a partial Schur form is constructed automatically without further reordering.

In the following, we compare TRQ and RKS on a 340×340 Tolosa matrix [2]. The Tolosa matrix is a model problem that has the important features of matrices that arise in the stability analysis of an airplane in flight. The full spectrum of this matrix is plotted in Figure 4.2. Eigenvalues with largest imaginary parts are of interest. We use the RKS code developed by Ruhe [16] for comparison. The same random starting vector is used in both RKS and TRQ. In the RKS code, Ritz values are computed

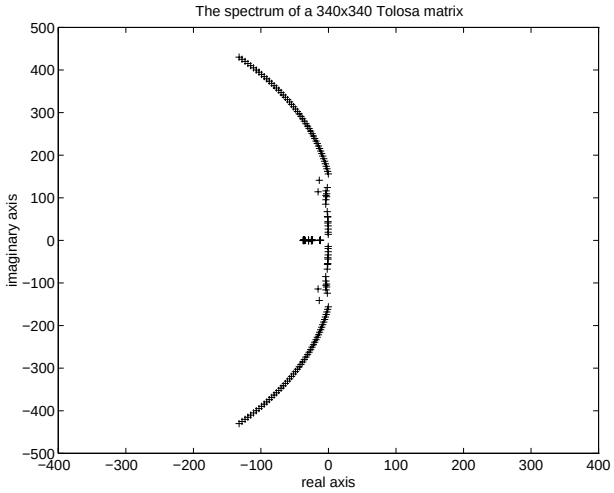


FIG. 4.2. *The spectrum of a* 340×340 *Tolosa matrix.*

TABLE 4.6
Comparison of IRA and TRQ on a Tolosa matrix.

Eigenvalue	RKS	RKS	TRQ(6)	TRQ(6)	TRQ(6)	IRA(10)
	$m = 5$	$m = 10$	$m = 1$	$m = 5$	$m = 10$	
$-132.3 + 430.1i$	12	14	11	14	15	14
$-127.9 + 425.2i$	15	16	13	16	17	51
$-123.5 + 420.2i$	17	18	15	17	18	64
$-119.3 + 415.2i$	19	20	17	19	20	125
$-115.1 + 410.2i$	31	22	19	21	22	201
Factorizations	9	6	19	7	6	1

from (4.1) at each iteration. A Ritz value is flagged as converged when the Ritz estimate falls below $tol = 10^{-10}$. The initial shift is placed at $\mu = -150 + 410i$. The same shift is used for at most m iterations. A new shift is selected after the current shift has been used for m iterations or after convergence of a Ritz value. The same shift selection strategy is used in TRQ for comparison. In Table 4.6, we list the first five computed eigenvalues and the number of iterations taken before each eigenvalue has converged. We choose $m = 5$ and $m = 10$ in RKS. The size of the Arnoldi factorization used in TRQ is set to be 6 ($k = 6$.) We tried $m = 1$ (optimal shift selection), $m = 5$, and $m = 10$ in TRQ. At the bottom of the table, we accumulated the total number of factorizations used in each run. For $m = 5$, the convergence history of RKS and TRQ are plotted in Figures 4.3 and 4.4, respectively. In these figures, we plot the residual norm of each approximate eigenvalue against the number of flops (floating point operations.) The vertical dotted line marks the end of each iteration; the dash-dot line marks the end of a matrix factorization. It is observed from Table 4.6 that it takes more than 10 iterations for both RKS and TRQ to locate the first eigenvalue. Once the first one emerges, both algorithms converge at a rate of two iterations per eigenvalue. Notice that horizontal axes in Figures 4.3 and 4.4 are labeled with different scales. For this problem, RKS builds a larger subspace than TRQ in order to capture all desired eigenpairs. Thus, more orthogonalizations are performed in RKS. This explains the larger number of flops required by RKS.

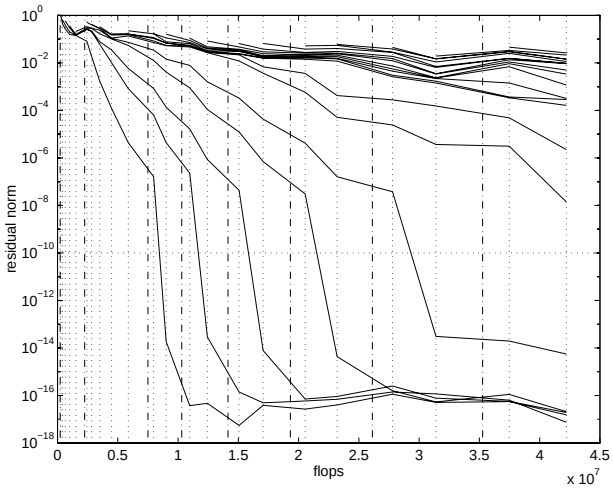


FIG. 4.3. *The convergence history of RKS.*

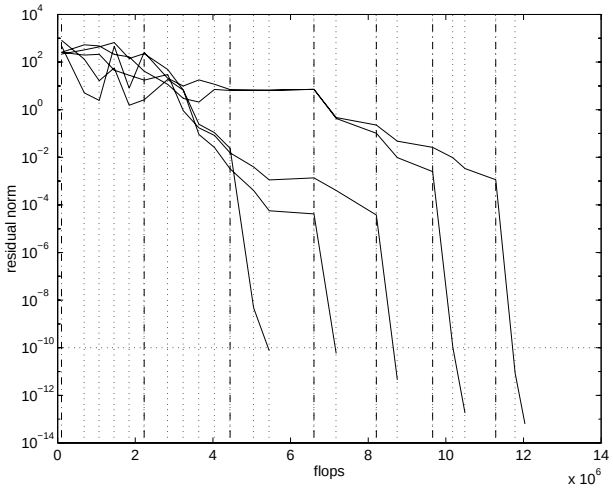


FIG. 4.4. *The convergence history of TRQ.*

Residual norms of all Ritz pairs are plotted in Figure 4.3. Only five of them have converged to the desired tolerance of 10^{-10}. Clearly, TRQ is competitive with RKS in terms of both the number of factorizations and the number of iterations, and both algorithms compare favorably with IRA with $p = 10$ (IRA(10)).

4.4. Comparison with JDQR. If factoring $A - \mu I$ is inexpensive, we may consider using an optimal shift described in section 3.2 in each TRQ iteration. In this case, the performance of TRQ is comparable with that of the Jacobi–Davidson method.

Given an initial approximation v_0 of a desired eigenvector, the Jacobi–Davidson method [19] finds, at each step, a correction vector z_k that is orthogonal to the previous approximate eigenvector u_k. A new subspace is created by adjoining this vector to the previous subspace and taking the span. The next approximate eigenpairs are drawn

from projection onto the new subspace. The correction vector z_k is obtained from the equation

$$(I - u_k u_k^H)(A - \theta_k I)(I - u_k u_k^H) z_k = -r_k \quad \text{and} \quad z_k \perp u_k, \tag{4.2}$$

where $r_k = Au_k - \theta_k u_k$ and θ_k is the current approximation to the eigenvalue of interest. It can be shown [19] that if (4.2) is solved exactly, the Jacobi–Davidson method becomes the accelerated inverse iteration, i.e., it builds an orthonormal basis of the subspace

$$\mathcal{S}(A, v_0, \{\theta_j\}) = \text{span}\{v_0, v_1, v_2, \ldots v_k\},$$

where $v_j = (A - \theta_j I)^{-1} v_{j-1}$. Ritz approximations are extracted from this subspace. It is shown in [16] that this method is equivalent to RKS with an optimal shift selected in each iteration. The subspace $\mathcal{S}(A, v_0, \{\theta_j\})$ is not a Krylov subspace. The Hessenberg relationship (3.1) is not preserved in the Jacobi–Davidson iteration. To obtain several eigenvalues and eigenvectors, some standard deflation schemes [17] are needed. To avoid building a large dimensional subspace $\mathcal{S}$, restarting is also necessary. The implementation of the JDQR algorithm is explained in detail in [9]. We compare the performance of TRQ and JDQR on a standard eigenvalue problem arising from the stability analysis of the Brusselator wave model (BWM) [2]. Eigenvalues with largest real parts are of interest. They help to determine the existence of stable periodic solutions to the Brusselator wave equation as some parameter varies. The size of the matrix we choose is 200×200. The 32 rightmost eigenvalues are plotted in Figure 4.5. We place the target shift at $\sigma = 1.0$ and use TRQ and JDQR to find 4 eigenvalues closest to σ. In Table 4.7, we list the first four computed eigenvalues and the number of factorizations used to obtain each one of them. In the runs using TRQ, we tried $k = 5$ and $k = 8$. In JDQR, the maximum dimension of subspace from which approximate eigenpairs are drawn is 8. Restart begins at the 6th column ($j_{min} = 5$.) It is denoted by JDQR(5,8) in Table 4.7.

It is observed from Table 4.7 that TRQ takes fewer iterations to find all four eigenvalues of interest. However, as pointed out in [9], the correction equation may

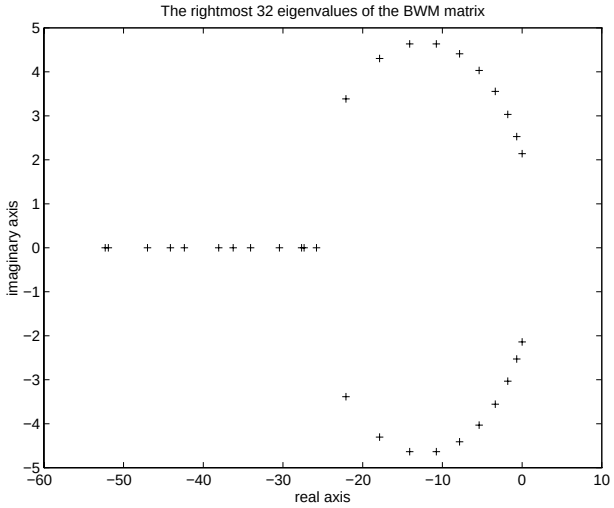


FIG. 4.5. *The* 32 *rightmost eigenvalues of a* 200×200 *BWM matrix.*

TABLE 4.7
Comparison of TRQ and JDQR on the BWM problem.

Eigenvalue	TRQ(5)	TRQ(8)	JDQR(5,8)
$1.820 \times 10^{-5} + 2.140i$	8	6	14
$1.820 \times 10^{-5} - 2.140i$	11	8	17
$-0.6747 + 2.529i$	14	12	19
$-0.6747 - 2.529i$	16	14	21

be solved by one step of GMRES iteration in the first j_{min} steps of JDQR iterations. This is equivalent to building the initial Jacobi–Davidson search space [9] by running a j_{min}-step Arnoldi iteration. For the BWM problem, this technique reduces the total number of exact solves in JDQR(5,8) to 16.

5. Inexact TRQ and restarting. Rapid convergence of the TRQ algorithm is observed in section 4 when the TRQ equation

$$(I - V_k V_k^H)(A - \mu I)(I - V_k V_k^H)v_+ = v\alpha \quad \text{with} \quad V_k^H v_+ = 0, \quad \|v_+\| = 1 \tag{5.1}$$

is solved exactly in each iteration. In this section, we explore the possibility of relaxing the solution accuracy of (5.1) while maintaining the rapid convergence of TRQ iteration. This is extremely important for many applications in which the factorization of $A - \mu I$ is too costly, and an approximate solution of $(A - \mu I)x = b$ can be provided by an iterative solver.

Recall that one of the important characteristics of the TRQ algorithm is the inverse iteration relation between the first column of V_k^+ and the first column of V_k, i.e.,

$$(A - \mu I)v_1^+ = v_1.$$

If an optimal shift is chosen at each iteration, the convergence of v_1 to an eigenvector of A is often quadratic or cubic. We will show in the following that if the projected equation is solved approximately, an inexact inverse iteration is maintained between v_1^+ and v_1. Superlinear convergence can still be achieved if optimal shifts are used.

Suppose $\tilde{v}_+$ is an approximate solution to (5.1). Since $I - V_k V_k^H$ is a projection, we may replace $\tilde{v}_+$ with $(I - V_k V_k^H)\tilde{v}_+$ in (5.1). Thus, we explicitly orthogonalize the approximate solution $\tilde{v}_+$ against all columns of V_k through

$$\tilde{v}_+ \leftarrow (I - V_k V_k^H)\tilde{v}_+$$

and normalize it so that $\|\tilde{v}_+\| = 1$. The unknowns h and α present in (2.4) are then computed directly as if $\tilde{v}_+$ were an exact solution to (5.1), i.e.,

$$\tilde{h} \leftarrow V_k^H(A - \mu I)\tilde{v}_+ = V_k^H A\tilde{v}_+, \quad \tilde{\alpha} \leftarrow v^H(A - \mu I)\tilde{v}_+.$$

These lead to the equation

$$(A - \mu I)(V_k, \tilde{v}_+) = (V_k, v)\begin{pmatrix} H_k - \mu I_k & \tilde{h} \\ \beta_k e_k^T & \tilde{\alpha} \end{pmatrix} + z e_{k+1}^T, \tag{5.2}$$

where $z e_{k+1}^T$ is an error term with

$$z \equiv (A - \mu I)\tilde{v}_+ - (V_k, v)\begin{pmatrix} \tilde{h} \\ \tilde{\alpha} \end{pmatrix}.$$

By construction, z satisfies

$$V_k^H z = 0, \quad v^H z = 0.$$

We may now compute an upper triangular $\hat{R} = \left(\begin{smallmatrix} R_k & r \\ 0 & \rho \end{smallmatrix}\right)$ and an orthogonal $\hat{Q} = \left(\begin{smallmatrix} Q_k & q \\ \sigma e_k^T & \gamma \end{smallmatrix}\right)$ such that

$$\begin{pmatrix} H_k - \mu I_k & \tilde{h} \\ \beta_k e_k^T & \tilde{\alpha} \end{pmatrix} = \hat{R}\hat{Q}$$

and multiply (5.2) from the right by $\hat{Q}^H$ to get

$$(A - \mu I)(V_k Q_k^H + \tilde{v}_+ q^H, v_k \bar{\sigma} + \tilde{v}_+ \bar{\gamma}) = (V_k, v) \begin{pmatrix} R_k & r \\ 0 & \rho \end{pmatrix} + z(q^H, \bar{\gamma}).$$

The first column of $V_k^+ = V_k Q_k^H + \tilde{v}_+ q^H$ is related to the first column of V_k through the equation

$$(A - \mu I) v_1^+ = \rho_{11} v_1 + z\delta, \tag{5.3}$$

where $v_1^+ = V_k^+ e_1$ and δ is the first element of the vector q. Since the orthogonal matrix $\hat{Q}$ is constructed from accumulation of a sequence of Givens rotations used in the RQ factorization, δ is a product of $(k-1)$ sines. Its magnitude is bounded by 1 and it is likely to be quite small due to the accumulated product of sines. Thus, the error term present in the inexact inverse iteration (5.3) is at worst of the same magnitude as the error introduced in solving (5.1) and is very likely to be much smaller. In fact if the first subdiagonal element β_1 is small (indicating the (1,1) element of H_k is nearly an eigenvalue of A), then $|\delta|$ is very likely to be smaller than $|\beta_1|$ which may be verified by considering the effect of the final Givens rotation to occur in the RQ step. Therefore, the error committed by accepting the inexact solution to the linear system (5.1) is damped by the RQ step to obtain a more accurate inverse-iteration relation between the vectors v_1^+ and v_1 than might be expected.

We would like to continue the TRQ update as described in Steps 2.4–2.6 of Algorithm 2. However, because of the error incurred in (5.2), the updated orthonormal basis $V_k^+ = V_k Q^H + \tilde{v}_+ q^H$ no longer spans a Krylov subspace. However, the first column of V_k^+ is approximately what we would have obtained if the TRQ equation is solved exactly. Thus, one may recover a truncated Hessenberg reduction by running a k-step Arnoldi process with v_1^+ as the starting vector. We refer to this step as a *restart*. The restarted TRQ (RTRQ) iteration is summarized in Algorithm 4 in Figure 5.1.

If a Krylov subspace type of method (such as conjugate gradient or GMRES) is used to solve the TRQ equation in step 2.2 of the above algorithm, it may be of advantage to work with the operator $B \equiv (I - V_k V_k^H)(A - \mu I)(I - V_k V_k^H)$ directly since B may be better conditioned in the subspace $V_k^\perp$. Of course, the matrix B need not be formed explicitly, only the matrix vector multiplication Bv is required.

5.1. Comparison with JDQR. In the following, we present a numerical example of using the inexact RTRQ to compute the eigenvalues of the CK656 matrix described in [2]. Eigenvalues of this matrix all have multiplicity two. We look for 4 eigenvalues near the target shift $\sigma = 5.0$, and set $k = 5$ in RTRQ (RTRQ(5)). The computational result is compared with JDQR with $j_{min} = 5$, $j_{max} = 8$ (JDQR(5,8)).

Algorithm 4: Truncated RQ iteration with restart (RTRQ)

Input: (A, V_k, H_k, f_k) with $AV_k = V_k H_k + f_k e_k^T, V_k^H V_k = I$, H_k upper Hessenberg.
Output: (V_k, H_k) such that $AV_k = VH_k, V_k^H V_k = I$ and H_k is upper triangular.

1. Put $\beta_k = \|f_k\|$ and put $v = f_k/\beta_k$;
2. **for** $j = 1, 2, 3, ...$ until *convergence*,
  2.1. Select a shift $\mu \leftarrow \mu_j$;
  2.2. Solve $(I - V_k V_k^H)(A - \mu I)(I - V_k V_k^H)w = v$ approximately;
  2.3. $w \leftarrow (I - V_k V_k^H)w$, $v_+ \leftarrow w/\|w\|$;
  2.4. $h \leftarrow V_k^H A v_+, \quad \alpha \leftarrow v^H (A - \mu I) v_+$;
  2.5. *RQ* Factor $\begin{pmatrix} H_k - \mu I_k & h \\ \beta_k e_k^T & \alpha \end{pmatrix} = \begin{pmatrix} R_k & r \\ 0 & \rho \end{pmatrix} \begin{pmatrix} Q_k & q \\ \sigma e_k^T & \gamma \end{pmatrix}$;
  2.6. $v_1 \leftarrow V_k Q_k^H e_1 + v_+ q^H e_1$;
  2.7. Restart: $(H_k, V_k, v, \beta_k) \leftarrow$ Arnoldi(A, v_1);
3. **end**;

FIG. 5.1. *Restarted TRQ iteration.*

TABLE 5.1
Comparison of RTRQ and JDQR on the CK656 problem.

Eigenvalue	RTRQ(5)	JDQR(5,8)
5.5024	3	15
5.5024	6	23
1.5940	11	25
1.5940	17	35

The same random starting vector is used in both tests. The TRQ equation and the projected correction equation in JDQR are solved by GMRES with no preconditioning or restart. The maximum GMRES steps allowed in each linear solve is set to be 10. The GMRES residual tolerance is set to be 10^{-6}. The optimal shift selection strategy is used in both tests; i.e., the Ritz value that is the nearest to the target shift but has not converged is used as the next shift. No tracking [9] is used in JDQR. In Table 5.1, we list the four eigenvalues of interest and the number of iterations taken by RTRQ and JDQR before each eigenvalue has converged. We observe that for this example, RTRQ takes fewer iterations than JDQR to capture eigenvalues of interest. In particular, RTRQ is able to capture the first eigenvalue much quicker than JDQR. However, RTRQ costs more per iteration than JDQR because the projection in the TRQ equation always involves k vectors, and k matrix vector multiplications must be performed in each iteration to reconstruct an Arnoldi factorization. Thus, the overall performance should be compared in terms of total number of matrix vector multiplications or flops used in both methods. This is illustrated in Figure 5.2. We plot the residual of each approximate eigenpair against the number of flops. The residuals of the approximate eigenpairs are monitored one at a time. When the residual curve corresponding to the approximation to the eigenpair (λ_j, z_j) drops below 10^{-9}, we start to monitor and record the residual for the next approximate eigenpair (λ_{j+1}, z_{j+1}). We should point out that the comparison made here is still preliminary.

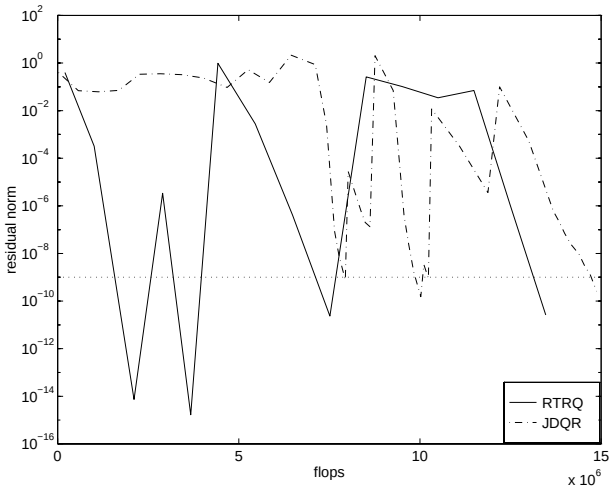


FIG. 5.2. *Convergence history of RTRQ and JDQR for the CK*656 *matrix.*

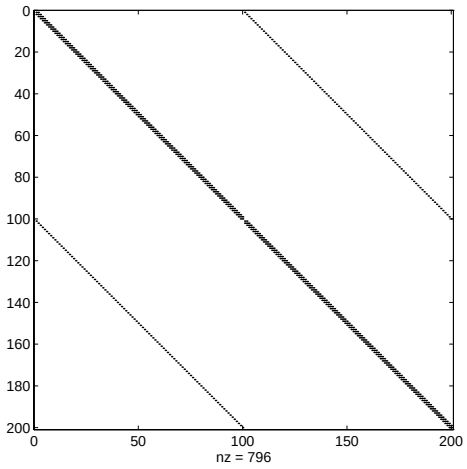


FIG. 5.3. *The structure of a* 200×200 *BWM matrix.*

Several techniques are available to improve the performance of JDQR [9] and many of these may be used in RTRQ as well.

5.2. The effect of preconditioning. Solving the TRQ equation is the most expensive part of the TRQ iteration. When an iterative method is used, a good preconditioner may accelerate the convergence and reduce the overall cost. The improved accuracy in the solution to the TRQ equation often brings about a reduction in the total number of TRQ iterations.

One may precondition the projected system

$$(I - V_k V_k^T)(A - \mu I)(I - V_k V_k^T)w = v$$

directly to obtain an approximate solution to the TRQ equation. However, it may not be easy to find a good preconditioner M for the projected matrix $(I - V_k V_k^T)(A -$

TABLE 5.2
Comparison of RTRQ with and without preconditioner.

Eigenvalue	Diagonal	ILU(0)	Tridiagonal
$1.820 \times 10^{-5} + 2.140i$	80	21	7
$1.820 \times 10^{-5} - 2.140i$	> 100	38	12
$-0.6747 + 2.529i$	> 100	52	19
$-0.6747 - 2.529i$	> 100	66	23

$\mu I)(I - V_k V_k^T)$. Instead, one usually has a preconditioner for the matrix A. As pointed out in [9], this preconditioner may need to be projected into $V_k^\perp$ in order to accelerate the convergence of the Jacobi–Davidson iteration. The projected shifted preconditioner is sometimes not a good preconditioner for the projected shifted matrix A. This extra projection does not seem to be necessary in the TRQ iteration since the TRQ equations may be solved using the scheme discussed in section 3. This scheme solves a linear system $(A - \mu I)w = v$. Thus, a preconditioner of A may be easily applied. In the following we present an example that demonstrates the effect of preconditioning on the restarted TRQ iteration. Four eigenvalues of the BWM matrix used in section 4 are computed, and the size of the Arnoldi factorization in the TRQ iteration is set to be 5 ($k = 5$.) The target shift is placed at 1.0. The TRQ equation is solved using a preconditioned GMRES with no restart. The maximum number of GMRES iterations allowed in each solve is set to be 10. The GMRES residual tolerance is set to be 10^{-6}. The structure of the BWM matrix is shown in Figure 5.3. We used the diagonal part, the tridiagonal part, and the incomplete LU factors (ILU(0)) of the matrix A as the preconditioner. The number of iterations used to obtain the four eigenvalues near 1.0 are listed in Table 5.2. Without a preconditioner no eigenvalue is found in 100 iterations. The convergence history of RTRQ with various preconditioners is shown in Figure 5.4. The residual norm of each approximate eigenpair is plotted against the number of flops subsequentially. The solid curve corresponds to RTRQ with tridiagonal preconditioning. The dashed curve corresponds to RTRQ with ILU(0) preconditioning. The dash-dot curve corresponds to RTRQ with diagonal preconditioning. The dotted curve is associated with RTRQ with no preconditioning. When the residual curve drops below the dotted line indicating the acceptable residual tolerance 10^{-9}, we start to monitor and record the residual of the next approximate eigenpair. It is observed that a good preconditioner improves the convergence of RTRQ dramatically.

5.3. Comparison with accelerated inverse iteration with Wielandt deflation. The inexact TRQ iteration with restart does not completely mimic the exact TRQ. In particular, the truncated Hessenberg reduction is enforced through an Arnoldi iteration rather than an implicit RQ update. The method behaves more like a single vector iteration with deflation than an RQ iteration in which the rapid convergence of one eigenvalue is often accompanied with the convergence of other eigenvalues at a slower pace.

In this section, we compare restarted TRQ with the accelerated inverse iteration combined with a deflation scheme that is very close to the Wielandt deflation (INVWD) [17, p. 117] for computing a few eigenvalues of A. We show that the exact TRQ performs better than the exact INVWD and the inexact TRQ appears to be more reliable than the inexact INVWD.

The inverse iteration can be viewed as a shifted and inverted power iteration. It

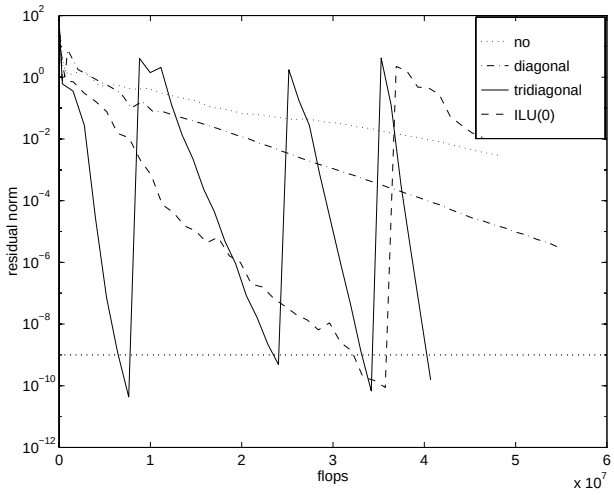


FIG. 5.4. *Convergence history of preconditioned TRQ for the BWM matrix.*

requires solving

$$(A - \mu I)w = v,$$

where v is the previous approximation to an eigenvector and w is the current approximation. The acceleration is achieved by choosing, at each iteration, a shift μ that is the best approximation to the desired eigenvalue. Once an eigenpair (λ, u) has been found, the next pair may be obtained by applying shifted power iteration to the deflated operator $A_1 = (A - \mu I)^{-1} - uq^H$, where $q = (A - \mu I)^{-H}u$. This deflation scheme is an variant of the *explicit* Wielandt deflation [23, p. 596], [17, p. 117]. The deflated operator $A_1 = (I - uu^H)(A - \mu I)^{-1}$ does not preserve right eigenvectors of A in general, unless A is normal. However, it does preserve Schur vectors of A. Thus, to generalize this deflation scheme for a converged invariant subspace, one should replace u with a matrix of Schur vectors U that spans the converged invariant subspace and satisfies $U^H U = I$. This is a more stable variant of a technique referred to as the Schur–Wielandt deflation in [17, p. 122]. It leads to the algorithm INVWD (Figure 5.5) which we adopt here for comparison to RTRQ.

In the following, we first present an example that demonstrates the advantage of using TRQ over using inverse iteration with Schur–Wielandt-like deflation. Then we compare the performance of the inexact TRQ with restart to the inverse iteration in which the linear system is solved approximately.

In the first example, we choose A to be the two-dimensional discrete Laplacian used before. Six eigenvalues of the smallest magnitude are computed. The size the Arnoldi factorization maintained in the TRQ iteration is 7 ($k = 7$.) The same size is chosen for the deflated Arnoldi iteration used in INVWD to help determine the shift. The same random starting vector is used in both TRQ and INVWD. In INVWD, a Ritz pair (μ_j, z_j) is considered to be converged if the direct residual norm $\|r_j\| = \|Az_j - \mu_j z_j\|$ falls below $tol = 10^{-12}$. In TRQ, the convergence criterion is a tolerance of machine epsilon in the test for declaring a subdiagonal element to zero. Table 5.3 shows the number of iterations taken before each eigenvalue has converged. In Figure 5.6, the convergence history of the residual for each computed eigenpair is

(INVWD) A Schur–Wielandt deflated inverse iteration

Input: (A, μ, v, U) such that (μ, v) is the current approximation to the desired eigenpair, and columns of U contain the converged Schur vectors.

Output: A new approximate eigenpair (μ_+, v_+) that may be used in the next cycle of an inverse iteration.

1. Solve $(A - \mu I)w = v$;
2. $v \leftarrow (I - UU^H)w$; $v \leftarrow v/\|v\|$;
3. $f \leftarrow Av$; $\alpha = v^H w$;
4. $H_1 = (\alpha)$; $V = (v)$; $f \leftarrow f - v\alpha$;
5. $f \leftarrow (I - UU^H)f$;
6. **for** $j = 1, 2, ..., k$
  6.1. $\beta_j = \|f\|$; $v_{j+1} \leftarrow f/\beta_{j+1}$;
  6.2. $V_{j+1} = (V_j,\ v_{j+1})$; $H_j \leftarrow \begin{pmatrix} H_j \\ \beta_j e_j^T \end{pmatrix}$;
  6.3. $z \leftarrow Av_{j+1}$; $z \leftarrow (I - UU^H)z$;
  6.4. $h \leftarrow V_j^H z$; $H_{j+1} = (H_j,\ h)$;
  6.5. $f \leftarrow z - V_{j+1}h$;
7. **end**;
8. Compute an desired Ritz pair (μ_+, v_+) from H_k and V_k to be used in the next cycle of an inverse iteration.

FIG. 5.5. *Schur–Wielandt deflated inverse iteration.*

shown. The height of each circle and star corresponds to the residual of the eigenpair computed by TRQ and INVWD, respectively. The TRQ residuals corresponding to the approximations to the same eigenpair are connected by a solid line. The INVWD residuals are connected by a dash-dot line. The circles below the dotted line correspond to the residuals of converged eigenpairs computed by TRQ. It is easily observed that the global convergence of TRQ is better than INVWD. In INVWD, every residual curve starts from the top ($\|r\| \approx 10^{-1}$), whereas in TRQ, the convergence of the second and fifth eigenpairs are followed by the immediate convergence of the third and the sixth pairs. The residual for the fifth eigenpair starts from roughly 10^{-10} and drops below 10^{-14} in one iteration. We should also mention that the convergence

TABLE 5.3
Comparison of TRQ and INVWD on a two-dimensional Laplacian.

Eigenvalue	TRQ	INVWD
0.16203	3	3
0.39851	6	7
0.39851	7	13
0.63499	10	17
0.77129	12	20
0.77129	13	24

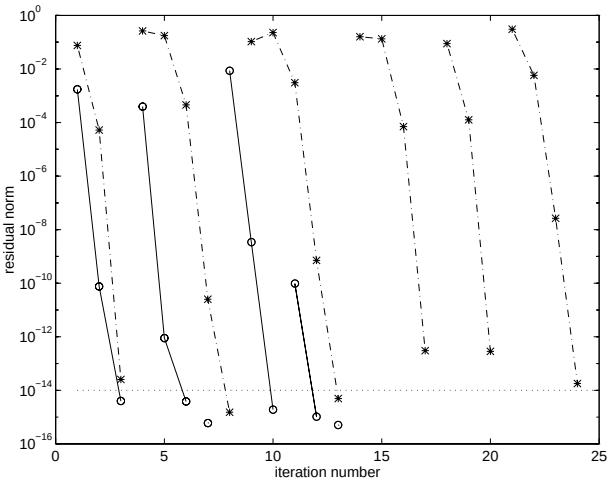


FIG. 5.6. *Traces of the residual in TRQ and INVWD.*

of INVWD is sensitive to the starting vector and the size of the subspace used to obtained the shift. Eigenvalues may not necessarily converge in order. For example, large eigenvalues may appear early when we look for the ones with the smallest magnitude.

In the next example, we compare the performance of the inexact TRQ with that of the inexact INVWD. We consider computing eigenvalues of the DW1024 matrix that arises from dielectric waveguide problems in integrated circuit applications [2]. Four eigenvalues near 1.0 are of interest. In both methods, linear systems are solved by GMRES with no restart. The maximum number of GMRES iterations allowed is set to be 10. The GMRES residual tolerance is set to be 10^{-8}. The size of the Arnoldi factorization maintained in the inexact TRQ iteration is set to be 5 ($k = 5$.) The same size is set for the deflated Arnoldi iteration used in INVWD to determine the shift. The traces of the residual for each computed eigenpair are shown in Figure 5.7. Residual norms are plotted against the number of flops. The solid curve corresponds to the residual norm of the inexact TRQ. The dotted curve corresponds to the residual of the inexact INVWD. We observe that the inexact INVWD converges much slower than the inexact TRQ.

6. Conclusions. This development of the TRQ iteration has led to a promising way to take advantage of situations when shift-invert equations can be solved directly and also when they can only be solved inexactly through iterative means. We have demonstrated with several numerical experiments that this scheme provides a promising and competitive alternative to rational Krylov methods and the Jacobi–Davidson method in the two respective cases. The scheme is relatively simple and very efficient in terms of required numerical computation compared to these and other related methods. Finally, the convergence properties and deflation schemes are easily understood through the close connection with the RQ iteration for dense matrices.

Future research will focus upon analyzing the filtering properties obtained from embedding the shift-invert equations in the TRQ iteration. Equation (5.3) indicates a damping of the error introduced by inexact solution when the RQ iteration is carried out. The numerical properties and implications of this phenomenon are not yet

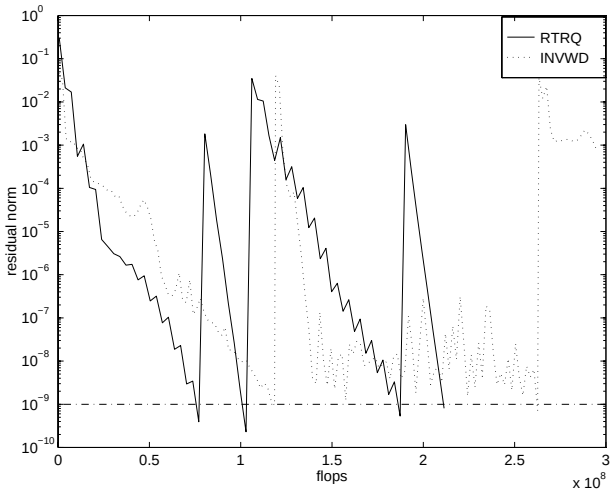


FIG. 5.7. *Traces of the residual in inexact TRQ and INVWD.*

understood.

We chose the GMRES method to solve the TRQ equation iteratively in the inexact TRQ method because of its simplicity and reliability. Certainly, other iterative solvers such as QMR, BICGSTAB could have been used. It would be interesting to compare the performance of these iterative solvers in the TRQ context. More research is required with respect to preconditioners and how they should be utilized within the TRQ equations. Exhaustive computational experimentation and comparisons are needed to determine whether the TRQ equations should be solved in bordered form, projected form, or by utilizing Lemma 3.1. These are issues both for direct and iterative solutions of the TRQ equations. The extension of these ideas to the generalized eigenvalue problem will also be important. Eventually, we expect to produce numerical software based upon this scheme to complement the IRA schemes already available in ARPACK.

Acknowledgments. We would like to thank R. B. Lehoucq and G. L. G. Sleijpen for reading the manuscript in detail and providing us with numerous corrections and suggestions. In particular, suggestions from Drs. Sleijpen and Lehoucq improved sections 3.1 and 5.3, respectively. We would also like to thank the anonymous referees for careful reading and helpful comments.

REFERENCES

[1] J. BAGLAMA, D. CALVETTI, AND L. REICHEL, *Iterative methods for the computation of a few eigenvalues of a large symmetric matrix*, BIT, 36 (1996), pp. 400–421.

[2] Z. BAI, R. BARRETT, D. DAY, J. DEMMEL, AND J. DONGARRA, *Test Matrix Collection* (*Non-Hermitian Eigenvalue Problems*), Research report, Department of Mathematics, University of Kentucky, Lexington, KY, 1995.

[3] Z. BAI AND G. W. STEWART, *SRRIT–A FORTRAN subroutine to calculate the dominant invariant subspace of a nonsymmetric matrix*, ACM Trans. Math. Software, to appear.

[4] T. BRACONNIER, *The Arnoldi–Tchebycheff Algorithm for Solving Large Nonsymmetric Eigenproblems*, Technical Report TR/PA/93/25, CERFACS, Toulouse, France, 1993.

[5] D. CALVETTI, L. REICHEL, AND D. C. SORENSEN, *An implicitly restarted Lanczos method for large symmetric eigenvalue problems*, ETNA, 2 (1994), pp. 1–21.

[6] F. Chatelin, *Eigenvalues of Matrices*, Wiley, New York, 1993.
[7] J. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram–Schmidt QR factorization*, Math. Comp., 30 (1976), pp. 772–795.
[8] I. S. Duff and J. A. Scott, *Computing selected eigenvalues of sparse unsymmetric matrices using subspace iteration*, ACM Trans. Math. Software, 19 (1993), pp. 137–159.
[9] D. R. Fokkema, G. L. G. Sleijpen, and H. A. Van der Vorst, *Jacobi–Davidson style QR and QZ algorithm for partial reduction of matrix pencils*, SIAM J. Sci. Comput., to appear.
[10] R. W. Freund and N. M. Nachtigal, *QMRPACK: A package of QMR algorithms*, ACM Trans. Math. Software, 22 (1996), pp. 46–77.
[11] R. B. Lehoucq and D. C. Sorensen, *Deflation techniques for an implicitly restarted Arnoldi iteration*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 789–821.
[12] R. B. Lehoucq, D. C. Sorensen, P. Vu, and C. Yang, *ARPACK: An Implementation of the Implicitly Restarted Arnoldi Iteration That Computes Some of the Eigenvalues and Eigenvectors of a Large Sparse Matrix*, 1995. Available from ftp.caam.rice.edu from the directory pub/software/ARPACK.
[13] R. B. Morgan, *On restarting the Arnoldi method for large nonsymmetric eigenvalue problems*, Math. Comp., 65 (1996), pp. 1213–1230.
[14] A. Ruhe, *The rational Krylov algorithm for nonsymmetric eigenvalue problems*, III: *Complex shifts for real matrices*, BIT, 34 (1994), pp. 165–176.
[15] A. Ruhe, *Rational Krylov algorithms for nonsymmetric eigenvalue problems*, II: *Matrix pairs*, Linear Algebra Appl., 197/198 (1994), pp. 283–295.
[16] A. Ruhe, *Rational Krylov, a Practical Algorithm for Large Sparse Nonsymmetric Matrix Pencils*, Technical Report UCB/CSD-95-871 (revised), Computer Science Division (EECS), University of California, Berkeley, CA, 1995.
[17] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halsted Press, New York, 1992.
[18] J. A. Scott, *An Arnoldi code for computing selected eigenvalues of sparse real unsymmetric matrices*, ACM Trans. Math. Soft., 21 (1995), pp. 432–475.
[19] G. L. G. Sleijpen and H. A. Van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.
[20] G. L. G. Sleijpen, J. G. L. Booten, D. R. Fokkema, and H. A. Van der Vorst, *Jacobi–Davidson type methods for generalized eigenproblems and polynomial eigenproblems*, BIT, 36 (1996), pp. 595–633.
[21] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.
[22] D. C. Sorensen and C. Yang, *A Truncated RQ-Iteration for Large Scale Eigenvalue Calculations*, Technical Report TR96-06, Department of Computational and Applied Mathematics, Rice Univeristy, Houston, TX, 1996.
[23] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, Oxford, UK, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 1074–1096, October 1998

015

BULGE EXCHANGES IN ALGORITHMS OF *QR* TYPE*

DAVID S. WATKINS†

Abstract. The QR algorithm and its variants are among the most popular methods for calculating eigenvalues of matrices. Typical implementations chase bulges from top to bottom of an upper Hessenberg matrix. It is also possible to chase bulges from bottom to top. There are some situations in which it may be advantageous to chase bulges in both directions at once, in which case one needs a procedure for passing bulges through each other without mixing up the information that the bulges convey. This paper derives a procedure for passing bulges of arbitrary degree through each other. Experiments with a Fortran 90 program show that the procedure works well in practice for bulges of degree two.

Key words. eigenvalue computation, QR algorithm, bulge exchanges

AMS subject classifications. 65F15, 15A18

PII. S0895479896299950

1. Introduction.

The family of QR-like algorithms [12], [7], [8], [15] is one of the most prominent classes of algorithms for solving matrix eigenvalue problems. Most implementations of QR-like algorithms are bulge chasing algorithms. The matrix is first transformed to a condensed form (usually upper Hessenberg) by a similarity transformation. Then each QR iteration consists of a sequence of similarity transformations, which first create a bulge in the condensed form, then chase the bulge from one end of the matrix to the other.

For simplicity we will confine ourselves in this paper to the upper Hessenberg condensed form. Typically, bulges are created at the top of the matrix and chased downward, but it is also sometimes useful to create bulges at the bottom of the matrix and chase them upward. One might also wish to create bulges at both ends of the matrix and chase them toward each other. Then one has to ask how to pass the bulges through each other without mixing up the information that they contain. This paper will address that problem.

The problem of passing bulges through each other, or bulge exchange, was first considered by Byers [2], [3], who showed how to exchange two bulges of degree one or two in a Hamiltonian Hessenberg matrix. The motive was to create a version of the QR algorithm that preserves the Hamiltonian form. This is achieved by requiring that the transforming matrices be symplectic, which requires in turn that two "mirror image" bulges be chased in opposite directions. Watkins [16] showed that the bulge exchange does not depend on Hamiltonian structure; it can be done in arbitrary upper Hessenberg matrices, not only for QR algorithms, but for LR, SR, and other QR-like algorithms as well. However, the discussion in [16] was restricted to bulges of degree one.

Obviously it must be possible to exchange bulges of arbitrary degree, but for years we could not see how to do it. Finally we were able to discern the mechanism by which

*Received by the editors March 4, 1996; accepted for publication (in revised form) by B. Kågström May 20, 1997; published electronically July 17, 1998.
http://www.siam.org/journals/simax/19-4/29995.html
†Department of Pure and Applied Mathematics, Washington State University, Pullman, WA 99164-3113 (watkins@wsu.edu). Current address: 6835 24th Avenue NE, Seattle, WA 98115-7037.
This research was supported by the National Science Foundation under grant DMS-9403569.

information is carried by bulges [18], and this proved to be the key to understanding the bulge exchange process in general.

To give an idea of where bulge exchanges may be useful beyond the preservation of Hamiltonian form, we shall sketch two potential applications. First, bidirectional chasing (chasing bulges in both directions) may prove to be an effective strategy for decreasing memory traffic in parallel QR codes on distributed memory machines. Blocks of the matrix are doled out to the processors. Whenever a bulge gets to a block boundary, information has to be passed back and forth between processors. If a bulge from above arrives at a block boundary at the same time as another bulge arrives from below (and this can certainly be arranged), then information about both bulges can be swapped with no more effort than it would take to pass information about a single bulge. Thus the total message passing is decreased substantially.

The second potential application is of a completely different nature. There is a class of algorithms, exemplified by the LR algorithm without pivoting [22], [5], that operate on highly condensed, e.g., tridiagonal, forms. Algorithms of this type are able to preserve the highly condensed form because they forego pivoting. The price they pay for this is that they can be unstable because they must sometimes use extremely large multipliers (i.e., small pivots) in the elimination operations that the algorithm performs. A typical safeguard against this is to save a copy of the matrix before each iteration. If at any point during the iteration a multiplier exceeds a certain tolerance, the iteration is aborted and restarted with different shifts. The ability to pass bulges through each other offers an alternative to the wasteful restarting process. If at any time during the bulge chase it is found that an excessive multiplier is about to arise, the bulge can simply be stopped. A new bulge chase, with different shifts, can be initiated, and the new bulge can be passed right through the old one, which waits in place until conditions improve. Once one or two (or perhaps 100) bulges have passed through, conditions will become favorable for chasing the original bulge forward.

This paper's contents. This paper derives a general procedure for passing bulges of arbitrary size through each other. Section 2 sets the stage by introducing explicit and implicit GR and RG algorithms and establishing the relationships between them. The implicit algorithms are the ones that chase bulges. A new explanation of the mechanism by which the bulges carry information is presented. Section 3, the heart of the paper, derives a procedure for exchanging bulges without mixing up the information that they carry. Section 4 describes numerical experiments that demonstrate the viability of the bulge exchange procedure for bulges of degree two in the QR (unitary) case.

Notation. The vector space of complex n-tuples is denoted $\mathbb{C}^n$. The standard basis vectors in $\mathbb{C}^n$ are denoted e_i, $i = 1, \ldots, n$. The ith entry of e_i is 1 and all other entries are 0. Technically we should write $e_i^{(n)}$ to specify the dimension of the space in which e_i lies, but we shall delete the n, which will always be clear from context. The space of $m \times n$ complex matrices will be denoted $\mathbb{C}^{m\times n}$. a_{ij} and $a_{ij}^{(k)}$ denote the (i,j) entries of matrices A and A_k, respectively. A^* is the matrix whose (i,j) entry is $\overline{a}_{ji}$. I_m is the identity matrix in $\mathbb{C}^{m\times m}$. Script letters such as $\mathcal{S}$ will be used to denote subspaces of $\mathbb{C}^n$. $\mathcal{S}^\perp$ denotes the orthogonal complement of $\mathcal{S}$ in $\mathbb{C}^n$.

2. Explicit and implicit GR algorithms.

***GR* and *RG* decompositions.** We will work within the context of the class of QR-like algorithms known as GR algorithms [20]. [1] The basic operation in an iteration of a GR algorithm is a GR decomposition. Given a matrix $B \in \mathbb{C}^{n\times n}$, a GR *decomposition* of B is a factorization of B into a product

$$B = GR,$$

where G is nonsingular and R is upper triangular. There are many ways to do a GR decomposition. For example, one can require that G be a unitary matrix, in which case the decomposition is called a QR decomposition. Every B has a QR decomposition; under mild conditions [2] the QR decomposition is essentially unique [15]. Another possibility is to require that G be lower triangular with ones on the main diagonal, that is, *unit* lower triangular. Then the decomposition is called an LR (or LU) decomposition. Almost every B has an LR decomposition, which is unique [15].

One can equally well perform an RG decomposition, $B = RG$, putting the upper triangular matrix first. There are many different types of RG decompositions, including RQ and RL.

Explicit *GR* and *RG* algorithms. Let us begin with GR algorithms. These are iterative processes; for our purposes it suffices to focus on a single iteration, starting with a matrix $A \in \mathbb{C}^{n\times n}$ and resulting in a matrix $\hat{A}$ that is similar to A. First a *spectral transformation function* f is chosen. For the purpose of specifying the class of algorithms, the only requirement on f is that the matrix $f(A)$ be well defined. The next step is to perform a GR decomposition: $f(A) = GR$. Just as there is much latitude in how f is chosen, there is also a great deal of choice in how the GR decomposition is carried out. Now there is the additional requirement that $f(A)$ have a GR decomposition of the desired type (e.g., LR). If this is not the case, then either f or the type of GR decomposition being used has to be changed. Once a GR decomposition has been performed, the final step is to use G to perform a similarity transformation on A; specifically $\hat{A} = G^{-1}AG$. The whole process is summarized by two equations

$$f(A) = GR, \qquad \hat{A} = G^{-1}AG.$$

We call this a GR iteration *driven by* f. If this process is performed repeatedly with intelligent choices of f and G, the iterates will tend to (block) upper triangular form, revealing the eigenvalues of A.

In addition to GR algorithms there are RG algorithms, which are defined in a similar fashion. An RG iteration driven by f is summarized by the two equations

$$f(A) = RG, \qquad \hat{A} = GAG^{-1}.$$

As we shall see, the RG algorithm is the GR algorithm with time reversal. Repeated application of the RG algorithm will also lead to (block) upper triangular form, but (assuming fixed f, for example) the (blocks of) eigenvalues will come out in the opposite order on the main diagonal.

[1] All of the developments of this paper have extensions that can be applied to GZ algorithms for the generalized eigenvalue problem [11], [8], [15], [21].

[2] For example, B is nonsingular or B is a proper upper Hessenberg matrix.

If A has spectrum $\lambda_1, \ldots, \lambda_n$, then $f(A)$ has spectrum $f(\lambda_1), \ldots, f(\lambda_n)$. The point of applying f is to spread out the spectrum. The more spread out it is (plotting magnitudes $|f(\lambda_i)|$ on a logarithmic scale), the faster the iterates converge [20].

Commonly f is taken to be a polynomial, but in this paper we will focus on rational f. Let us suppose to begin with that $f = 1/q$, where q is a polynomial whose zeros are not eigenvalues of A. A GR iteration driven by $1/q$ has the form

$$q(A)^{-1} = GR, \qquad \hat{A} = G^{-1}AG.$$

Notice that this is the same as an RG iteration driven by q, since the equations can be rewritten as

$$q(A) = \tilde{R}\tilde{G}, \qquad \hat{A} = \tilde{G}A\tilde{G}^{-1},$$

where $\tilde{G} = G^{-1}$ and $\tilde{R} = R^{-1}$.

Now suppose $f = p/q$, where p and q are polynomials. A GR iteration has the form

$$q(A)^{-1}p(A) = GR, \qquad \hat{A} = G^{-1}AG. \tag{2.1}$$

This can be decomposed into two iterations, a GR iteration driven by p followed by an RG iteration driven by q, as follows. First

$$p(A) = G_1R_1, \qquad \check{A} = G_1^{-1}AG_1.$$

Then

$$q(\check{A}) = R_2G_2, \qquad \hat{A} = G_2\check{A}G_2^{-1}.$$

Using the equation $q(A) = G_1q(\check{A})G_1^{-1} = G_1R_2G_2G_1^{-1}$, we see easily that $\hat{A}$ and A are related by (2.1), where $G = G_1G_2^{-1}$ is nonsingular and $R = R_2^{-1}R_1$ is upper triangular. Thus a GR iteration driven by a rational function can be accomplished by a GR iteration driven by the numerator polynomial followed by an RG iteration driven by the denominator polynomial. One easily checks that the same result is achieved if the RG step is done first; the order of the steps is immaterial.

Remark. Nothing that has been said so far depends on p and q being polynomials. Consider a GR step driven by any function p for which $p(A)$ is invertible. Follow that with an RG step driven by the same p. As we have just seen, the result is a GR step driven by $p/p = 1$. Now $1(A) = I$, whose spectrum is as unspread as possible; a GR step driven by I goes nowhere. This shows that the RG iteration driven by p cancels the GR iteration driven by p. In other words, an RG iteration is a GR iteration run in reverse.

Implicit *GR* and *RG* algorithms. If p is a polynomial, even if its degree is as low as two, computing $p(A)$ is an expensive proposition. Therefore GR iterations are usually effected by a process that avoids the explicit computation of $p(A)$. These are called *implicit* GR iterations, and they require that the matrix first be reduced to a condensed form such as upper Hessenberg. Let us assume, therefore, from this point on that A is a proper upper Hessenberg matrix. *Upper Hessenberg* means almost upper triangular: $a_{ij} = 0$ for $i > j + 1$. *Proper* means that $a_{i,i-1} \neq 0$ for $i = 2, \ldots, n$.

The key results are the following theorems. Notice that if $p(A) = GR$, then the first column of G is proportional to $p(A)e_1$, which is the first column of $p(A)$. Theorem 2.1 is a sort of converse of this simple observation.

THEOREM 2.1. *Let $A \in \mathbb{C}^{n\times n}$ be a proper upper Hessenberg matrix, and let p be a polynomial. Let G be a nonsingular matrix whose first column is proportional to $x = p(A)e_1$, such that $\hat{A} = G^{-1}AG$ is upper Hessenberg. Then there exists an upper triangular matrix R such that $p(A) = GR$.*

This is Theorem 2.4 of [19]. It says that if we can manage a similarity transformation that has the right first column and preserves upper Hessenberg form, we will have done a GR iteration. The next theorem is the RG analogue of Theorem 2.1.

THEOREM 2.2. *Let $A \in \mathbb{C}^{n\times n}$ be a proper upper Hessenberg matrix, and let q be a polynomial. Let G be a nonsingular matrix whose last row is proportional to $y^T = e_n^T q(A)$, such that $\hat{A} = GAG^{-1}$ is upper Hessenberg. Then there exists an upper triangular matrix R such that $q(A) = RG$.*

Thus if we can manage a similarity transformation that has the right last row and preserves upper Hessenberg form, we will have done an RG iteration. Theorems 2.1 and 2.2 are special cases of Theorems 2.9 and 2.10, which will be proved below.

Similarity transformations satisfying the conditions of Theorems 2.1 and 2.2 can be effected by bulge chasing algorithms. Chasing algorithms that perform GR iterations were described in [19]. These algorithms begin by doing a similarity transformation that creates a bulge in the upper Hessenberg form at the top. Then they return the matrix to upper Hessenberg form by similarity transformations that clear the columns one by one, proceeding from left to right. The effect of this is to chase the bulge down the main diagonal and off of the bottom edge of the matrix.

For variety's sake we will describe here bulge chasing algorithms that perform RG iterations implicitly. These create a bulge at the bottom and chase it to the top. Suppose we wish to carry out an RG iteration driven by the polynomial q of degree k. According to Theorem 2.2, we need a similarity transformation whose last row is proportional to $y^T = e_n^T q(A)$, the last row of $q(A)$. We wish to avoid calculating $q(A)$ explicitly. It turns out that computing only the last row is much cheaper if $k \ll n$, as we shall assume. The polynomial q is normally given in factored form $q(z) = (z-\tau_1)(z-\tau_2)\cdots(z-\tau_k)$. Thus

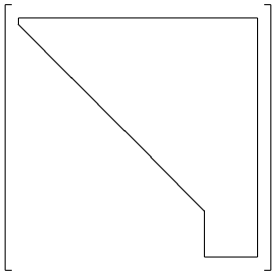
$$y^T = e_n^T(A-\tau_1 I)\cdots(A-\tau_k I),$$

which can be computed by a sequence of k matrix vector multiplications. Since e_n has only its last entry nonzero and A is upper Hessenberg, only the bottom $k+1$ rows of A are involved in the computation, only the last $k+1$ entries of y^T can be nonzero, and y^T can be computed in $O(k^3)$ flops. We have $y_i = 0$ for $i = 1, \ldots, n-k-1$. It is important to note that $y_{n-k} \neq 0$, because A is properly upper Hessenberg. Indeed $y_{n-k} = a_{n,n-1}a_{n-1,n-2}\cdots a_{n-k+1,n-k}$.

Let $G_0 = \operatorname{diag}\{I_{n-k-1}, \tilde{G}_0\}$, where $\tilde{G}_0$ is a matrix whose last row is proportional to $\tilde{y}_0^T = \begin{bmatrix} y_{n-k}\cdots y_n \end{bmatrix}$. Thus the last row of G_0 is proportional to y^T.

Clearly $e_{k+1}^T\tilde{G}_0 = \alpha_0^{-1}\tilde{y}_0^T$ for some nonzero α_0. Another way to put this is to say that $\tilde{G}_0$ is chosen in such a way that $\tilde{G}_0^{-1}$ "eliminates" or "clears out" $\tilde{y}_0^T$ in the sense that $\tilde{y}_0^T\tilde{G}_0^{-1} = \alpha_0 e_{k+1}^T$.

Let $A_1 = G_0 A G_0^{-1}$. This matrix is not upper Hessenberg; it has a bulge at the



bottom that protrudes k diagonals beyond the subdiagonal. The tip is at the position $a^{(1)}_{n,n-k-1}$. This entry is certainly nonzero. In fact $a^{(1)}_{n,n-k-1} = \alpha_0^{-1} y_{n-k} a_{n-k,n-k-1}$. We call this a bulge of *degree* k.

The rest of the iteration consists of a sequence of similarity transformations that return the matrix to upper Hessenberg form by clearing out the rows one by one from bottom to top. The first step is to choose a G_1 such that G_1^{-1} clears out the last row of A_1. The form of G_1 is $\operatorname{diag}\{I_{n-k-2}, \tilde{G}_1, 1\}$, where the last row of $\tilde{G}_1$ is proportional to

$$\tilde{y}_1^T = \begin{bmatrix} a^{(1)}_{n,n-k-1} & \cdots & a^{(1)}_{n,n-1} \end{bmatrix}.$$

Thus $e_{k+1}^T \tilde{G}_1 = \alpha_1^{-1} \tilde{y}_1^T$ or $\tilde{y}_1^T \tilde{G}_1^{-1} = \alpha_1 e_{k+1}^T$. When A_1 is tranformed to $A_1 G_1^{-1}$, the last row is returned to upper Hessenberg form; the bulge is reduced by one row. When the similarity transformation is completed by left multiplication by G_1, a column is added to the left-hand edge of the bulge. In other words, the matrix $A_2 = G_1 A_1 G_1^{-1}$ has a bulge whose tip is at $a^{(2)}_{n-1,n-k-2}$. This tip entry is certainly nonzero. Indeed $a^{(2)}_{n-1,n-k-2} = \alpha_1^{-1} a^{(1)}_{n,n-k-1} a_{n-k-1,n-k-2}$.

This step establishes the pattern for the process. The next transforming matrix G_2 has the form $\operatorname{diag}\{I_{n-k-3}, \tilde{G}_2, I_2\}$, where $\tilde{G}_2$ is chosen so that $\tilde{G}_2^{-1}$ returns row $n-1$ to Hessenberg form and so on. After $n-1$ steps the iteration is complete. Taking $\hat{A} = A_{n-1}$, we have $\hat{A} = GAG^{-1}$, where $G = G_{n-2} \cdots G_1 G_0$. In this matrix product every factor except G_0 has the general form $\operatorname{diag}\{M, 1\}$. Consequently the last row of G is the same as the last row of G_0, which is proportional to the last row of $q(A)$. Since $\hat{A}$ is upper Hessenberg, we conclude from Theorem 2.2 that the bulge chasing algorithm affects an iteration of an RG algorithm driven by q.

This development establishes a somewhat weak connection between explicit and implicit versions of RG algorithms. A much more detailed connection is established in [21].

The relationship between y and q. It is obvious that q determines y uniquely (given a fixed A) through the equation $y^T = e_n^T q(A)$. It is worth noting that y determines q uniquely as well.

PROPOSITION 2.3. *Let A be a properly upper Hessenberg matrix. Let $y \in \mathbb{C}^n$. Then there is a unique polynomial q of degree less than n such that $y^T = e_n^T q(A)$. If $y_i = 0$ for $i = 1, \ldots, n-k-1$ and $y_{n-k} \neq 0$, then q has degree exactly k.*

Proof. Since A is properly upper Hessenberg, the row vectors e_n^T, $e_n^T A$, $e_n^T A^2$, $\ldots$, $e_n^T A^{n-1}$ are linearly independent. Indeed, the first nonzero entry of $e_n^T A^k$ is in position $n-k$. Thus there exist unique coefficients $c_0, \ldots, c_{n-1}$ such that $y^T = e_n^T (c_0 I + c_1 A + c_2 A^2 + \cdots + c_{n-1} A^{n-1})$. The proposition follows easily. ∎

This "row-wise" result has the following column-wise analogue.

PROPOSITION 2.4. *Let A be a properly upper Hessenberg matrix. Let $z \in \mathbb{C}^n$. Then there is a unique polynomial q of degree less than n such that $z = q(A)e_1$. If $z_i = 0$ for $i = k+2, \ldots, n$ and $z_{k+1} \neq 0$, then q has degree exactly k.*

How information is carried in the bulge. In an RG or GR iteration driven by the polynomial q, the information that needs to be carried in the bulge is the polynomial q itself, which we shall assume is monic, without loss of generality. q can be factorized completely over the complex field: $q(z) = (z - \tau_1) \cdots (z - \tau_k)$, and it is usually presented in this form. Knowledge of the roots $\tau_1, \ldots, \tau_k$, which are called the *shifts* for the iteration, is equivalent in principle to knowledge of q.

How is this information carried in the bulge? We already answered this question in [18]. Here we shall prove the main results of [18] by a different approach, getting some new insights along the way.

A good overall view of the bulge chasing process is achieved by embedding the pencil $A - \mu I_n$ in a larger pencil $\tilde{A} - \mu N_{n+1}$ obtained by adjoining a column on the left and a row at the bottom.

$$\tilde{A} - \mu N_{n+1} = \begin{bmatrix} x & A - \mu I_n \\ 0 & y^T \end{bmatrix}. \tag{2.2}$$

We will consider various choices of vectors x and y. Notice that N_{n+1} is not an identity matrix; it is a strictly upper triangular matrix with ones on the superdiagonal. The symbol N stands for nilpotent. Since N_{n+1} is singular, a pencil of this form must have at least one infinite eigenvalue.

Let us first consider the case $x = e_1$, $y = e_n$. Then the pencil $\tilde{A} - \mu N_{n+1}$ is upper triangular. Its determinant is the product of the subdiagonal entries of A, which is nonzero because A is properly upper Hessenberg. Thus the characteristic polynomial $\det(\tilde{A} - \mu N_{n+1})$ is a nonzero constant, which means that the pencil is regular and has no finite eigenvalues; all $n+1$ eigenvalues are infinite. In fact each main diagonal entry $a_{j+1,j} - 0\mu$ signals an infinite eigenvalue.

For the purpose of analyzing an iteration of the RG algorithm driven by q, we take

$$x = e_1 \quad \text{and} \quad y^T = e_n^T q(A). \tag{2.3}$$

Now the eigenvalues of (2.2) are not all infinite. The pencil is not upper triangular, but it is still block triangular. It has the form

$$\begin{bmatrix} H_0 & * \\ 0 & B_0 \end{bmatrix} - \mu N_{n+1}, \tag{2.4}$$

where H_0 is upper triangular and nonsingular, and $B_0 \in \mathbb{C}^{(k+1)\times(k+1)}$ would be upper triangular, except that its last row consists of $\tilde{y}^T$, the nonzero part of y^T. The asterisk denotes a submatrix whose values are not of immediate interest. The spectrum of $\tilde{A} - \mu N_{m+1}$ is the union of the spectra of $H_0 - \mu N_{n-k}$ and $B_0 - \mu N_{k+1}$. The spectrum of $H_0 - \mu N_{n-k}$ consists of ∞ repeated $n - k$ times. The spectrum of $B_0 - \mu N_{k+1}$ is much more interesting.

THEOREM 2.5. *If $q(z) = (z - \tau_1) \cdots (z - \tau_k)$, then the eigenvalues of $B_0 - \mu N_{k+1}$ are $\tau_1, \ldots, \tau_k, \infty$. In other words, the characteristic polynomial of $B_0 - \mu N_{k+1}$ is q.*

Proof. The entry in the lower left-hand corner of B_0 is y_{n-k}, which is nonzero. It follows easily that the characteristic polynomial $\det(B_0 - \mu N_{k+1})$ has degree k. Thus

the pencil is regular and has k finite eigenvalues and one infinite eigenvalue. We just need to show that the finite eigenvalues are $\tau_1, \ldots, \tau_k$.

Let τ_i be any one of the roots of q. Then $q(z) = r(z)(z - \tau_i)$ for some polynomial r of degree $k-1$. Thus $y^T = s^T(A - \tau_i I_n)$, where $s^T = e_n^T r(A)$. Only the last k entries of s can be nonzero; let $\tilde{s} \in \mathbb{C}^k$ denote this subvector. Then, in view of the definition of B_0, the equation $y^T = s^T(A - \tau_i I_n)$ is equivalent to

$$\begin{bmatrix} \tilde{s}^T & -1 \end{bmatrix} (B_0 - \tau_i N_{k+1}) = 0,$$

so τ_i is an eigenvalue of $B_0 - \mu N_{k+1}$ with left eigenvector $\begin{bmatrix} \tilde{s}^T & -1 \end{bmatrix}$.

The proof is now complete if $\tau_1, \ldots, \tau_k$ are distinct, as there can be no more than k finite eigenvalues. We can deduce the general result by continuity; just perturb the roots slightly to make them distinct. This causes a slight perturbation of $\tilde{y}$. Then invoke the fact that the roots of the characteristic polynomial depend continuously on the data.[3] □

Theorem 2.5 has an analogue for GR iterations, which we shall note for future reference. This is Theorem 1 of [18].

THEOREM 2.6. *Let q be a polynomial of degree k. Let $\tilde{A} - \mu N_{n+1}$ be given by (2.2), where $x = q(A)e_1$ and $y = e_n$. Let $B_0 - \mu N_{k+1}$ be the $(k+1) \times (k+1)$ subpencil from the upper left-hand corner of the pencil $\tilde{A} - \mu N_{n+1}$. Then the characteristic polynomial of $B_0 - \mu N_{k+1}$ is q.*

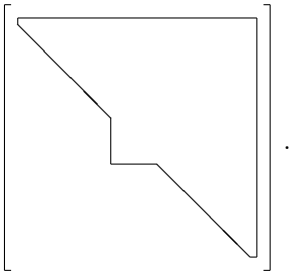
Theorem 2.5 implies that the eigenvalues of $\tilde{A} - \mu N_{n+1}$ given by (2.2) and (2.3) are ∞, repeated $n-k+1$ times, and $\tau_1, \ldots, \tau_k$. Now let us consider what happens to this pencil during an RG iteration that transforms A to $\hat{A}$. Each transformation that is applied to A has a corresponding transformation on the pencil. Under each such transformation the new pencil is equivalent to the old one, so the pencil's eigenvalues are preserved.

Consider the very first transformations. The transformation of A to AG_0^{-1} corresponds to multiplying the pencil $\tilde{A} - \mu N_{n+1}$ by $\mathrm{diag}\{1, G_0^{-1}\}$ on the right. This begins to form the bulge in A. Note, however, that the pencil already has a bulge: the nonzeros in the vector y disturb the triangular form; the part of the block named B_0 in (2.4) can be viewed as a bulge. Recall that the initial transformation G_0 is designed so that $y^T G_0^{-1} = \alpha_0 e_n^T$. This implies that when $\mathrm{diag}\{1, G_0^{-1}\}$ is applied to the pencil on the right, it clears out the last row of the bulge. The transformation from AG_0^{-1} to $A_1 = G_0 A G_0^{-1}$ corresponds to left multiplication of the pencil by $\mathrm{diag}\{G_0, 1\}$, and it adds one new column to the bulge.

This point of view shows that the initial similarity transformation, which we had viewed at first as a bulge-creating transformation, can also be viewed as a bulge chasing transformation.

[3] Another approach is to reverse the argument given here. If μ is an eigenvalue, then it has a left eigenvector, whose last entry is easily seen to be nonzero. This implies a relationship $y^T = s^T(A - \mu I_n)$ for some vector s with zeros in all but the last k positions. It follows that $q(z) = r(z)(z - \mu)$ for some r, so μ is one of the zeros of q.

At intermediate stages in the bulge chase the matrix A_i has the form



The corresponding pencil, which we can call $\tilde{A}_i - \mu N_{n+1}$, has a corresponding bulge which prevents it from being upper triangular. It is, however, block upper triangular. It can be written in the block form

$$\tilde{A}_i - \mu N_{n+1} = \begin{bmatrix} H_i - \mu N_j & * & * \\ & B_i - \mu N_{k+1} & * \\ & & K_i - \mu N_{n-k-j} \end{bmatrix},$$

where j is the column index of the first column of the bulge in A_i. $H_i - \mu N_j$ and $K_i - \mu N_{n-k-j}$ are upper triangular and have only ∞ as eigenvalues, repeated j and $n-k-j$ times, respectively. The pencil $B_i - \mu N_{k+1}$ contains the bulge, so let us call it the *bulge pencil*. Since $\tilde{A}_i - \mu N_{n+1}$ has the same eigenvalues as the original pencil $\tilde{A} - \mu N_{n+1}$, we can conclude that the eigenvalues of the bulge pencil are $\tau_1, \ldots, \tau_k$ and ∞. This shows how information about q is carried in the bulge; it is our main result.

THEOREM 2.7. *Consider a bulge chase driven by $q(z) = (z - \tau_1) \cdots (z - \tau_k)$. For $i = 1, 2, \ldots$ the k finite eigenvalues of the bulge pencil $B_i - \mu N_{k+1}$ are the shifts $\tau_1, \ldots, \tau_k$. In other words, the characteristic polynomial of the bulge pencil is q.*

Remark. This is a theoretical result, valid when the arithmetic is performed exactly. In floating-point arithmetic it can fail badly if k is large [18]. Thus one should use fairly small values of k (e.g., 2–6) in practice.

The local viewpoint. We have proved Theorem 2.7 by looking at the eigenvalues of the big pencil. This is the global viewpoint. A second approach (used in [18]) is to prove the result by induction. We know that the initial bulge pencil $B_0 - \mu N_{k+1}$ has eigenvalues $\tau_1, \ldots, \tau_k, \infty$. We can prove Theorem 2.7 by showing that $B_{i+1} - \mu N_{k+1}$ has the same eigenvalues as $B_i - \mu N_{k+1}$. This is the local viewpoint, and it has an important advantage. When we consider matrices with two or more bulges, we want to be sure that each bulge retains the shifts that it is carrying. We would not want shifts somehow to hop from one bulge to another. The local viewpoint shows that this does not happen; the shifts travel with the bulge.

The transformations that generate $B_{i+1} - \mu N_{k+1}$ from $B_i - \mu N_{k+1}$ are associated with the similarity transformation $A_{i+1} = G_i A_i G_i^{-1}$. Right multiplication of A_i by G_i^{-1} corresponds to an equivalence transformation on $B_i - \mu N_{k+1}$ that clears out the last row. This deflates the infinite eigenvalue. That is, the transformed pencil has block triangular form; the bottom 1×1 block has eigenvalue ∞, and the top $k \times k$ block has eigenvalues $\tau_1, \ldots, \tau_k$. Now delete the infinite eigenvalue by discarding the last row and column from the pencil. Next enlarge the pencil by adjoining a row and a column (obtained from the big pencil) at the top. This adds a new infinite eigenvalue. Now, left multiplication by G_i transforms this pencil to $B_{i+1} - \mu N_{k+1}$. This is an

equivalence tranformation, so it preserves the eigenvalues. Thus the eigenvalues of $B_{i+1} - \mu N_{k+1}$ are the same as those of $B_i - \mu N_{k+1}$.

The final configuration. At the end of the iteration, A has been transformed to $\hat{A} = GAG^{-1}$, and the pencil has been transformed to

$$\begin{bmatrix} z & \hat{A} \\ 0 & \alpha e_n^T \end{bmatrix} - \mu N_{n+1},$$

where $z = Ge_1$. The bulge has been chased from $\tilde{A}$, but the pencil still has a bulge, which has been compressed into z. Only the first $k+1$ entries of z can be nonzero, so we can express the pencil in block triangular form as

$$\begin{bmatrix} B_{n-1} - \mu N_{k+1} & * \\ & H_{n-1} - \mu N_{n-k} \end{bmatrix},$$

where $H_{n-1} - \mu N_{n-k}$ is upper triangular and has an $(n-k)$-fold eigenvalue ∞. $B_{n-1} - \mu N_{k+1}$ is the final bulge pencil of the RG iteration, and its characteristic polynomial is, of course, q. This is exactly the form of a pencil for the start of a GR iteration. If $\hat{A}$ is properly upper Hessenberg, we can perform a GR iteration that undoes the RG iteration by chasing the bulge back down to the bottom.

$\hat{A}$ will be properly upper Hessenberg as long as none of the shifts is an eigenvalue of A.[4] This is the generic case. If, on the other hand, some of the shifts are eigenvalues (i.e., $q(A)$ is singular), there will be a deflation at the end of the RG iteration [19], and the iteration will not be reversible.

We elaborate on these last remarks (generic case). Assuming $\hat{A}$ is properly upper Hessenberg, we can invoke Proposition 2.4 with A replaced by $\hat{A}$ to conclude that $z = \beta\tilde{q}(\hat{A})e_1$ for some unique monic polynomial $\tilde{q}$ of degree k and nonzero constant β. Theorem 2.6, applied with A, x, and q replaced by $\hat{A}$, z, and $\tilde{q}$, respectively, then implies that the characteristic polynomial of $B_{n-1} - \mu N_{k+1}$ is $\tilde{q}$. However, we already know that the characteristic polynomial is q. Thus $z = \beta q(\hat{A})e_1$, which is exactly the configuration we want for the beginning of a GR iteration on $\hat{A}$ [19].

Chasing bulges in both directions. Our discussion of explicit GR algorithms showed that a GR iteration driven by a rational function p/q can be broken into a GR iteration driven by the polynomial p and an RG iteration driven by the polynomial q. These can be performed in either order.

If we want to implement the process implicitly, our task is to chase a bulge from the top of the matrix to the bottom (GR iteration), then chase a different bulge from bottom to top (RG iteration). Alternatively we can do the upward chase before the downward chase. We can also consider chasing both bulges at once, which brings us to the heart of the paper.

If we wish to perform a GR iteration driven by p, we need to start by calculating $x = \alpha p(A)e_1$, which is used to create a bulge at the top of the matrix. Similarly, an RG iteration is started by calculating $y^T = \beta e_n^T q(A)$, which is used to create a bulge at the bottom. We can simultaneously perform similarity transformations that create both of these bulges and then push them toward each other. The upper (lower) bulge

[4]Indeed, in this case $q(A)$ is nonsingular, $R = q(A)G^{-1}$ is nonsingular, and $\hat{A} = R^{-1}AR$. The product of a properly upper Hessenberg matrix with two nonsingular upper triangular matrices is clearly properly upper Hessenberg.

has a bulge pencil whose characteristic polynomial is p (resp., q). In the big pencil viewpoint, we start with the configuration (2.2), where

$$x = \alpha p(A)e_1 \quad \text{and} \quad y^T = \beta e_n^T q(A).$$

In this configuration the pencil $\tilde{A} - \mu N_{n+1}$ has eigenvalues $\sigma_1, \ldots, \sigma_m$ and ∞ corresponding to the top bulge, $\tau_1, \ldots, \tau_k$ and ∞ corresponding to the bottom bulge, and ∞ with multiplicity $n-m-k-1$ in the middle.

Once the bulges have met, we need to pass them through each other somehow. The subpencil that contains the two bulges has the form

$$\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} - \mu N_{m+k+2},$$

where $B_{11} - \mu N_{m+1}$ and $B_{22} - \mu N_{k+1}$ have characteristic polynomials p and q, respectively. The challenge is to perform similarity transformations on the matrix that transform this pencil into the form

$$\begin{bmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{bmatrix} - \mu N_{m+k+2},$$

where $C_{11} - \mu N_{k+1}$ and $C_{22} - \mu N_{m+1}$ have characteristic polynomials q and p, respectively. That is, we want to interchange the positions of the shifts. If we can accomplish this, we can then complete the iteration by chasing the new bulges away from each other until they are pushed off of the ends of the matrix. Call the resulting matrix $\hat{A} = G^{-1}AG$. The final pencil is

$$\begin{bmatrix} z & \hat{A} \\ 0 & w^T \end{bmatrix} - \mu N_{n+1},$$

where z has nonzeros in at most its first $k+1$ positions, and w has nonzeros in at most its last $m+1$ positions. In block triangular form the pencil looks like

$$\begin{bmatrix} D - \mu N_{k+1} & * & * \\ & E - \mu N_{n-k-m-1} & * \\ & & F - \mu N_{m+1} \end{bmatrix}.$$

$D - \mu N_{k+1}$ and $F - \mu N_{m+1}$ have characteristic polynomials q and p, respectively.

From what we now know about the transmission of information in bulges, this procedure ought to effect a GR iteration driven by the rational function p/q. However, we have not yet proved that it does. We will prove it in the generic case when $\hat{A}$ is properly upper Hessenberg.[5]

Let us assume $\hat{A}$ is properly upper Hessenberg. Then, since $D - \mu N_{k+1}$ has q as its characteristic polynomial, we have $z = \gamma q(\hat{A})e_1$ by Proposition 2.4 and Theorem 2.6. Similarly, since $F - \mu N_{m+1}$ has p as its characteristic polynomial, $w^T = \delta e_n^T p(\hat{A})$ by Proposition 2.3 and Theorem 2.5. Summarizing the initial and final conditions, we have

$$x = \alpha p(A)e_1, \qquad y^T = \beta e_n^T q(A), \tag{2.5}$$

[5]Nothing bad happens in the nongeneric case. Indeed, deflations are highly desirable. However, the nongeneric case is harder to describe, and we prefer not to deal with the complications here.

$$z = \gamma q(\hat{A})e_1, \qquad w^T = \delta e_n^T p(\hat{A}). \tag{2.6}$$

These are clearly redundant; the conditions on x and z imply the conditions on y and w, for example. This follows from the invariance of the eigenvalues of the big pencil. The similarity transformation $\hat{A} = G^{-1}AG$ corresponds to the transformation

$$\begin{bmatrix} z & \hat{A} \\ 0 & w^T \end{bmatrix} = \begin{bmatrix} G^{-1} & \\ & 1 \end{bmatrix} \begin{bmatrix} x & A \\ 0 & y^T \end{bmatrix} \begin{bmatrix} 1 & \\ & G \end{bmatrix} \tag{2.7}$$

on the big pencil. Expanding this equation, we find that it is equivalent to the three equations

$$\hat{A} = G^{-1}AG, \qquad x = Gz, \qquad w^T = y^T G. \tag{2.8}$$

Theorem 2.9, which shows that the transformation $\hat{A} = G^{-1}AG$ is a GR iteration driven by p/q, will use the first two equations from (2.8) and the first equation from (2.5) and (2.6).

The proof of Theorem 2.9 depends on some basic facts about Krylov matrices. Given any $A \in \mathbb{C}^{n\times n}$ and any $v \in \mathbb{C}^n$, the *Krylov matrix* $K(A, v)$ is the $n \times n$ matrix whose columns are $v, Av, A^2v, \ldots, A^{n-1}v$. One easily proves the following proposition.

PROPOSITION 2.8. *If A is upper Hessenberg, then $K(A, e_1)$ is upper triangular. If A is properly upper Hessenberg, then $K(A, e_1)$ is nonsingular.*

THEOREM 2.9. *Suppose A and $\hat{A}$ are proper upper Hessenberg matrices satisfying $\hat{A} = G^{-1}AG$. Let $x = \alpha p(A)e_1$ and $z = \gamma q(\hat{A})e_1$ for some nonzero scalars α and γ, and suppose $x = Gz$. Then there exists upper triangular R such that*

$$q(A)^{-1}p(A) = GR.$$

Thus the similarity transformation $\hat{A} = G^{-1}AG$ is an iteration of the GR algorithm driven by the rational function p/q.

Proof. Since $G\hat{A} = AG$, we have $G\hat{A}^i = A^iG$ for all i, and $Gq(\hat{A}) = q(A)G$. Furthermore, the various hypotheses imply

$$\alpha p(A)e_1 = x = Gz = \gamma Gq(\hat{A})e_1 = \gamma q(A)Ge_1,$$

so $p(A)e_1 = \rho q(A)Ge_1$, where $\rho = \gamma/\alpha \neq 0$. More generally,

$$p(A)(A^ie_1) = \rho A^i q(A)Ge_1 = \rho q(A)G(\hat{A}^i e_1),$$

for $i = 0, 1, 2, \ldots, n-1$. This can be rewritten as

$$p(A)K(A, e_1) = \rho q(A)GK(\hat{A}, e_1).$$

Thus $q(A)^{-1}p(A) = GR$, where $R = \rho K(\hat{A}, e_1)K(A, e_1)^{-1}$ is upper triangular by Proposition 2.8. Because we are restricted to the generic case, $q(A)$ is guaranteed to be nonsingular [19]. ∎

If we take $q = 1$ in Theorem 2.9, we have $z = \gamma e_1$, and the theorem reduces to Theorem 2.1. The proof is valid even in the nongeneric case, because now $q(A) = I$. In the nongeneric case both $p(A)$ and $K(\hat{A}, e_1)$ are singular, but the proof remains valid.

Theorem 2.9 has a companion.

THEOREM 2.10. *Suppose A and $\hat{A}$ are proper upper Hessenberg matrices satisfying $\hat{A} = GAG^{-1}$. Let $y^T = \beta e_n^T q(A)$ and $w^T = \delta e_n^T p(\hat{A})$ for some nonzero scalars β and δ, and suppose $w^T G = y^T$. Then there exists upper triangular R such that*

$$q(A)p(A)^{-1} = RG.$$

Thus the similarity transformation $\hat{A} = GAG^{-1}$ is an iteration of the RG algorithm driven by the rational function q/p.

This is just a restatement of Theorem 2.9 with the symbols G and G^{-1} interchanged; the hypotheses of the two theorems are equivalent. If one wishes to prove Theorem 2.10 directly, one can do so with the help of Krylov matrices built by stacking rows of the form $e_n^T, e_n^T A, e_n^T A^2, \ldots$ from bottom to top.

If we take $p = 1$ in Theorem 2.10, we have $w^T = \delta e_n^T$, and the theorem reduces to Theorem 2.2. The proof is valid even in the nongeneric case.

3. Bulge exchanges. The following important question remains. Given a pencil

$$\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} - \mu N_{m+k+2}, \tag{3.1}$$

where $B_{11} - \mu N_{m+1}$ and $B_{22} - \mu N_{k+1}$ have characteristic polynomials p and q, respectively, how do we transform it to the form

$$\begin{bmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{bmatrix} - \mu N_{m+k+2}, \tag{3.2}$$

where $C_{11} - \mu N_{k+1}$ and $C_{22} - \mu N_{m+1}$ have characteristic polynomials q and p, respectively?

This is a block swapping problem similar to that of swapping two blocks of a matrix in real Schur form [1] or two blocks of a matrix pencil in generalized real Schur form [9], [10]. These swaps are accomplished by applying a transformation determined by solving a Sylvester or generalized Sylvester equation, respectively.

The method of [9], [10] cannot be applied directly to the pencil (3.1) because that method is designed for pencils $F - \lambda G$ with arbitrary G; it will not preserve the right-hand matrix N_{m+k+2}. In order to preserve the special form of (3.1) and (3.2), we must restrict ourselves to equivalence transformations of a certain type, namely, those that correspond to similarity transformations on the matrix in which B is embedded. Thus every application of a transformation $\mathrm{diag}\{1, G^{-1}\}$ on the right must be matched by a transformation $\mathrm{diag}\{G, 1\}$ on the left. We shall see how to transform (3.1) to (3.2) using only transformations of this type by solving a Sylvester-like equation.

A critical assumption will be that $\sigma_i \neq \tau_j$ for all i and j, where $\sigma_1, \ldots, \sigma_m$ and $\tau_1, \ldots, \tau_k$ are, as before, the zeros of p and q, respectively. This is the normal situation. After all, there is no profit in sending the same shift in opposite directions. For rapid convergence the τ's should be well separated from the σ's.

Deflating subspaces. We review some of the basic ideas associated with deflations of matrix pencils [13], [14, p. 752]. Let $B - \mu M$ be an $n \times n$ regular matrix pencil. *Regular* means that the characteristic polynomial $\det(B - \mu M)$ is not identically zero. Let $\mathcal{S}$ be a subspace of $\mathbb{C}^n$ of dimension j, with $0 < j < n$. Let $\mathcal{T}_B = B\mathcal{S} = \{Bx \mid x \in \mathcal{S}\}$, let $\mathcal{T}_M = M\mathcal{S}$, and let $\tilde{\mathcal{T}} = \mathcal{T}_B + \mathcal{T}_M$. Then $(B - \mu M)\mathcal{S} \subseteq \tilde{\mathcal{T}}$ for all μ. $\mathcal{S}$ is called a *deflating subspace* for the pencil $B - \mu M$ if $\dim(\tilde{\mathcal{T}}) \leq j$. If $\mathcal{S}$ is a deflating subspace and $\mathcal{T}$ is any j-dimensional subspace of $\mathbb{C}^n$ containing $\tilde{\mathcal{T}}$, then $(\mathcal{S}, \mathcal{T})$ is called a *deflating pair* of subspaces for the pencil $B - \mu M$.

The following characterization is clearly true. A pair $(\mathcal{S},\mathcal{T})$ of j-dimensional subspaces is a deflating pair if and only if $B\mathcal{S}\subseteq\mathcal{T}$ and $M\mathcal{S}\subseteq\mathcal{T}$.

There is also a useful matrix characterization of deflating pairs. Let $s_1,\dots,s_j$ and $t_1,\dots,t_j$ be bases of $\mathcal{S}$ and $\mathcal{T}$, respectively, and let $S_1=[s_1,\dots,s_j]\in\mathbb{C}^{n\times j}$ and $T_1=[t_1\cdots t_j]\in\mathbb{C}^{n\times j}$. Then $(\mathcal{S},\mathcal{T})$ is a deflating pair for the pencil $B-\mu M$ if and only if there exist matrices $B'_{11},M'_{11}\in\mathbb{C}^{j\times j}$ such that

$$BS_1=T_1B'_{11}\quad\text{and}\quad MS_1=T_1M'_{11}. \tag{3.3}$$

If equations (3.3) hold, then every eigenvalue of $B'_{11}-\mu M'_{11}$ is an eigenvalue of $B-\mu M$. Indeed, if $(B'_{11}-\nu M'_{11})v=0$, then $(B-\nu M)(S_1v)=0$. Notice that the eigenvector S_1v is a member of $\mathcal{S}$. Conversely, every eigenvector of $B-\mu M$ that belongs to $\mathcal{S}$ is associated with an eigenvector of $B'_{11}-\mu M'_{11}$. The j eigenvalues of $B'_{11}-\mu M'_{11}$ are called the eigenvalues of $B-\mu M$ *associated with* the deflating subspace $\mathcal{S}$.[6]

If it happens that $M\mathcal{S}=\mathcal{T}$, the bases can be chosen so that $Ms_i=t_i$, $i=1,\dots,j$. With this choice the equations (3.3) take the simpler form

$$BS_1=T_1F\quad\text{and}\quad MS_1=T_1,$$

which can also be written as the single equation

$$BS_1=MS_1F. \tag{3.4}$$

Then the eigenvalues of the pencil associated with $\mathcal{S}$ are just the eigenvalues of the matrix F. The equation $M\mathcal{S}=\mathcal{T}$ holds iff the matrix M'_{11} is nonsingular iff all of the eigenvalues of $B-\mu M$ associated with $\mathcal{S}$ are finite.

Knowledge of deflating subspaces allows one to transform the pencil to block triangular form. Let S_1 and T_1 be matrices satisfying (3.3), and let $S_2,T_2\in\mathbb{C}^{n\times(n-j)}$ be matrices chosen so that $S=\begin{bmatrix} S_1 & S_2\end{bmatrix}$ and $T=\begin{bmatrix} T_1 & T_2\end{bmatrix}$ are nonsingular. Let $B'-\mu M'=T^{-1}(B-\mu M)S$. Then (3.3) implies $(B-\mu M)S_1=T_1(B'_{11}-\mu M'_{11})$, from which we find that $B'-\mu M'$ has the form

$$\begin{bmatrix} B'_{11}-\mu M'_{11} & B'_{12}-\mu M'_{12}\\ 0 & B'_{22}-\mu M'_{22}\end{bmatrix}.$$

A pencil that has been transformed to this form has $(\mathcal{E}_j,\mathcal{E}_j)$ as a deflating pair, where $\mathcal{E}_j=\operatorname{span}\{e_1,\dots,e_j\}$.

Deflating subspaces are also known as *right deflating subspaces*. A *left deflating subspace* for $B-\mu M$ is one that is right deflating for $B^*-\bar\mu M^*$. Thus $(\mathcal{S},\mathcal{T})$ is a left deflating pair for $B-\mu M$ if and only if $B^*\mathcal{S}\subseteq\mathcal{T}$ and $M^*\mathcal{S}\subseteq\mathcal{T}$. The matrix characterization takes the form

$$S_1^*B=B'_{11}T_1^*\quad\text{and}\quad S_1^*M=M'_{11}T_1^*.$$

PROPOSITION 3.1. *The pair $(\mathcal{S},\mathcal{T})$ is right deflating for $B-\mu M$ if and only if $(\mathcal{T}^\perp,\mathcal{S}^\perp)$ is a left deflating pair. If $(\mathcal{S},\mathcal{T})$ is a deflating pair, then the sets of eigenvalues associated with $(\mathcal{S},\mathcal{T})$ and $(\mathcal{T}^\perp,\mathcal{S}^\perp)$ are complementary subsets of the spectrum of $B-\mu M$, counting multiplicity.*

Proof. $B\mathcal{S}\subseteq\mathcal{T}$ if and only if $B^*\mathcal{T}^\perp\subseteq\mathcal{S}^\perp$. This is true for any operator, so it is true for M also. The first part follows immediately.

[6] These values clearly do not depend on the choice of $\mathcal{T}$, nor on the choice of bases $s_1,\dots,s_j$ and $t_1,\dots,t_j$.

The second part is perhaps most easily seen by looking at matrices. If $(\mathcal{S},\mathcal{T})$ is a deflating pair, there is a similarity transformation to block triangular form

$$T^{-1}(B-\mu M)S = B'-\mu M' = \begin{bmatrix} B'_{11}-\mu M'_{11} & B'_{12}-\mu M'_{12} \\ 0 & B'_{22}-\mu M'_{22} \end{bmatrix}, \tag{3.5}$$

where $S = \begin{bmatrix} S_1 & S_2 \end{bmatrix}$ and $T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}$. The eigenvalues associated with $(\mathcal{S},\mathcal{T})$ are the eigenvalues of $B'_{11}-\mu M'_{11}$. It clearly suffices to prove that the eigenvalues associated with $(\mathcal{T}^\perp,\mathcal{S}^\perp)$ are the eigenvalues of $B'_{22}-\mu M'_{22}$. The similarity transformation (3.5) can also be written as

$$T^{-1}(B-\mu M) = (B'-\mu M')S^{-1}. \tag{3.6}$$

Define matrices $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}$ and $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$ by $U^* = S^{-1}$ and $V^* = T^{-1}$. Then $U^*S = I$, which implies that the columns of U_2 span $\mathcal{S}^\perp$. Similarly the columns of V_2 span $\mathcal{T}^\perp$. Writing (3.6) in block form, we have

$$\begin{bmatrix} V_1^* \\ V_2^* \end{bmatrix}(B-\mu M) = \begin{bmatrix} B'_{11}-\mu M'_{11} & B'_{12}-\mu M'_{12} \\ 0 & B'_{22}-\mu M'_{22} \end{bmatrix}\begin{bmatrix} U_1^* \\ U_2^* \end{bmatrix},$$

which implies

$$V_2^*B = B'_{22}U_2^* \quad \text{and} \quad V_2^*M = M'_{22}U_2^*.$$

This shows that $(\mathcal{T}^\perp,\mathcal{S}^\perp)$ is a left deflating pair (reproving the first part) whose associated eigenvalues are the eigenvalues of $B'_{22}-\mu M'_{22}$. ∎

Deflating subspaces in the context of bulge exchange. In the context of interest here, we have two bulges side by side. The pencil has the form (3.1). $(\mathcal{E}_{m+1},\mathcal{E}_{m+1})$ and $(\mathcal{E}^\perp_{m+1},\mathcal{E}^\perp_{m+1})$ are right and left deflating pairs associated with eigenvalues $\sigma_1,\dots,\sigma_m,\infty$, and $\tau_1,\dots,\tau_k,\infty$, respectively. We would like to reverse this configuration; we want a pencil of the form (3.2), where $(\mathcal{E}_{k+1},\mathcal{E}_{k+1})$ is the right deflating pair associated with $\tau_1,\dots,\tau_k,\infty$.

The bulge exchange procedure. We begin with the form (3.1). $B_{22}-\mu N_{k+1}$ has eigenvalues $\tau_1,\dots,\tau_k,\infty$. We want to move these eigenvalues to the top, so to speak. Our procedure begins by separating the infinite eigenvalue from the others. This yields the deflating subspace associated with the finite eigenvalues. For now we are working within the small pencil $B_{22}-\mu N_{k+1}$. The left deflating subspace associated with the eigenvalue ∞ is the left nullspace of N_{k+1}, which is $\mathcal{U} = \operatorname{span}\{e_{k+1}\}$, since $e_{k+1}^*N_{k+1} = 0$. Let v^* denote the last row of B_{22}. Then $e_{k+1}^*B_{22} = v^*$, so if we let $\mathcal{V} = \operatorname{span}\{v\}$, then $(\mathcal{U},\mathcal{V})$ is the left deflating pair for the eigenvalue ∞. By Proposition 3.1 the pair $(\mathcal{V}^\perp,\mathcal{U}^\perp)$ is the right deflating pair for the finite eigenvalues. Thus we just have to find $\operatorname{span}\{v\}^\perp$. Let U be a nonsingular matrix such that

$$v^*U = \alpha e_{k+1}^* \tag{3.7}$$

for some α, and let Y denote the $(k+1)\times k$ submatrix consisting of the first k columns of U. Then the columns of Y form a basis of $\mathcal{V}^\perp = \operatorname{span}\{v\}^\perp$.

Let $Z = N_{k+1}Y$. Then, since all of the eigenvalues associated with $\mathcal{V}^\perp$ are finite, the columns of Z are linearly independent and form a basis of $\mathcal{U}^\perp$. Therefore, as in (3.4), there is a $k\times k$ matrix F such that

$$B_{22}Y = ZF = N_{k+1}YF. \tag{3.8}$$

The eigenvalues of F are $\tau_1, \ldots, \tau_k$.

It is easy to compute F. Both $B_{22}Y$ and Z have bottom rows consisting entirely of zeros. Let C and E be the square matrices obtained by deleting these zero rows from $B_{22}Y$ and Z, respectively. Then $C = EF$. Since Z has full rank, E is nonsingular. Therefore $F = E^{-1}C$.

Now let us shift our attention to the larger pencil

$$\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} - \mu \begin{bmatrix} N_{11} & N_{12} \\ 0 & N_{22} \end{bmatrix},$$

where $N_{11} = N_{m+1}$ and $N_{22} = N_{k+1}$. We wish to find the right deflating subspace associated with $\tau_1, \ldots, \tau_k$. This is tantamount to satisfying a larger version of (3.8), which can also be written as $B_{22}Y = N_{22}YF$. Thus we seek $X \in \mathbb{C}^{(m+1)\times k}$ such that

$$\begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} N_{11} & N_{12} \\ 0 & N_{22} \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} F. \tag{3.9}$$

PROPOSITION 3.2. *Equation* (3.9) *has a unique solution* X *if* $\sigma_i \neq \tau_j$ *for all* i *and* j.

Proof. The second of the two blocks in equation (3.9) is just (3.8), so X just has to be chosen so that the first block is satisfied. This can be written as a Sylvester-like equation

$$B_{11}X - N_{11}XF = N_{12}YF - B_{12}Y. \tag{3.10}$$

Letting $Z = N_{11}X + N_{12}Y$, we can write this as the generalized Sylvester equation

$$B_{11}X - ZF = -B_{12}Y,$$

$$N_{11}X - ZI = -N_{12}Y.$$

This has a unique solution (X, Z), because the pencils $B_{11} - \mu N_{11}$ and $F - \mu I$ have disjoint spectra [4]. ∎

Our program is to transform

$$\begin{bmatrix} X \\ Y \end{bmatrix},$$

whose columns span the deflating subspace associated with $\tau_1, \ldots, \tau_k$, to the leading part of the pencil. We begin by deflating an infinite eigenvalue, pushing the two bulges together. This is achieved by right multiplication by $\mathrm{diag}\{I_{m+1}, U\}$ and left multiplication by $\mathrm{diag}\{I_m, U^{-1}, 1\}$. This is the same U as before, the one that satisfies $v^*U = \alpha e_{k+1}^*$. Let $\hat{B} = \mathrm{diag}\{I_m, U^{-1}, 1\}B\mathrm{diag}\{I_{m+1}, U\}$. Multiplying (3.9) on the left by $\mathrm{diag}\{I_m, U^{-1}, 1\}$ and making appropriate insertions, we obtain

$$\hat{B} \begin{bmatrix} X \\ \hat{I} \end{bmatrix} = N_{m+k+2} \begin{bmatrix} X \\ \hat{I} \end{bmatrix} F,$$

where $\hat{I} = U^{-1}Y = [e_1, \ldots, e_k] \in \mathbb{C}^{(k+1)\times k}$. By the construction of U, $\hat{B}$ has the block triangular form

$$\hat{B} = \begin{bmatrix} \tilde{B} & * \\ 0 & \alpha \end{bmatrix}.$$

If we drop the last row and column, we deflate an infinite eigenvalue. Doing so we obtain

$$\tilde{B}\begin{bmatrix} X \\ I_k \end{bmatrix} = N_{m+k+1}\begin{bmatrix} X \\ I_k \end{bmatrix} F. \tag{3.11}$$

Equation (3.11) shows that the columns of $\left[\begin{smallmatrix} X \\ I_k \end{smallmatrix}\right]$ span the right deflating subspace of $\tilde{B} - \mu N_{m+k+1}$ associated with $\tau_1, \ldots, \tau_k$. We want to move this space to the top of the pencil. This requires a transformation that compresses the nonzeros in $\left[\begin{smallmatrix} X \\ I_k \end{smallmatrix}\right]$ to the top. Bearing in mind the constraint on the form of the transformations that we are allowed to perform on $\tilde{B}$, we see that we will not be able to touch the first row of X. Denoting this row by u^T, we can write

$$\begin{bmatrix} X \\ I_k \end{bmatrix} = \begin{bmatrix} u^T \\ W \end{bmatrix}.$$

Let G be a nonsingular matrix such that

$$W = GR, \tag{3.12}$$

where $R \in \mathbb{C}^{(m+k)\times k}$ is upper triangular. From the form of W it is clear that R has full rank (k).

Let $\tilde{C} = \operatorname{diag}\{G^{-1}, 1\}\tilde{B}\operatorname{diag}\{1, G\}$. We are going to transform (3.11) into an equation involving $\tilde{C}$, but first let us note that

$$\begin{bmatrix} 1 & \\ & G^{-1} \end{bmatrix}\begin{bmatrix} u^T \\ W \end{bmatrix} = \begin{bmatrix} u^T \\ R \end{bmatrix}.$$

The bottom m rows of R are zero. In order to emphasize this fact we write

$$\begin{bmatrix} u^T \\ R \end{bmatrix} = \begin{bmatrix} \hat{R} \\ 0 \end{bmatrix},$$

where $\hat{R} \in \mathbb{C}^{(k+1)\times k}$ has full rank. Multiplying (3.11) on the left by $\operatorname{diag}\{G^{-1}, 1\}$ and making appropriate insertions, we obtain

$$\tilde{C}\begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} = N_{m+k+1}\begin{bmatrix} \hat{R} \\ 0 \end{bmatrix} F. \tag{3.13}$$

$\tilde{C}$ is not block triangular, but, as we shall see, it is nearly so. Before we can break out two separate bulges, we need to put back the infinite eigenvalue that we removed earlier. Before we do that, let us investigate the special properties of $\tilde{C}$.

First let $\tilde{C}_{NW}$ denote the $(k+1)\times(k+1)$ submatrix in the northwest corner of $\tilde{C}$. Then the top $k+1$ rows of (3.13) imply that

$$\tilde{C}_{NW}\hat{R} = N_{k+1}\hat{R}F, \tag{3.14}$$

which implies that the k-dimensional subspace spanned by the columns of $\hat{R}$ is a deflating subspace of the pencil $\tilde{C}_{NW} - \mu N_{k+1}$ corresponding to eigenvalues $\tau_1, \ldots, \tau_k$, the eigenvalues of F.

This assumes that $\tilde{C}_{NW} - \mu N_{k+1}$ is a regular pencil. Actually we cannot rule out the possibility that $\tilde{C}_{NW} - \mu N_{k+1}$ is singular; it can happen that the last row of $\tilde{C}_{NW}$ is identically zero. However, (3.14) holds in any event.

Now let $\tilde{C}_{SW}$ denote the $(m+1)\times(k+1)$ submatrix in the southwest corner of $\tilde{C}$. Notice that $\tilde{C}_{SW}$ and $\tilde{C}_{NW}$ overlap by one row. The bottom $m+1$ rows of equation (3.13) imply that $\tilde{C}_{SW}\hat{R}=0$. Thus all rows of $\tilde{C}_{SW}$ (which are members of $\mathbb{C}^{k+1}$) are orthogonal to the k linearly independent columns of $\hat{R}$, so they all lie in a one-dimensional subspace. Thus $\tilde{C}_{SW}$ has rank at most one. In fact the rank is exactly one; if it were zero, then $\tilde{C}-\mu N_{k+m+1}$ would be a singular pencil.

Remark. Let $\tilde{C}_{SE}$ be the $(m+1)\times(m+1)$ submatrix in the southeast corner of $\tilde{C}$ (overlapping with $\tilde{C}_{SW}$ by one column). One can show that the pencil $\tilde{C}_{SE}-\mu N_{m+1}$ is either singular (with its first column identically zero) or has $\sigma_1,\ldots,\sigma_m,\infty$ as its eigenvalues. This fact is not needed for our development, so we omit the proof.

One can also show that the pencil $\tilde{B}-\mu N_{m+k+1}$ has similar structure.

We can readily push the bulges apart again by applying a transformation that condenses the rank-one submatrix $\tilde{C}_{SW}$. Specifically, add one row and one column to the pencil $\tilde{C}-\mu N_{m+k+1}$ to obtain

$$\begin{bmatrix} \tilde{C} & * \\ 0 & c \end{bmatrix} - \mu N_{m+k+2}. \tag{3.15}$$

This restores the infinite eigenvalue that was deleted earlier. Let $x\in\mathbb{C}^{m+1}$ be a nonzero vector that is proportional to the columns of $\tilde{C}_{SW}$, and let V be a nonsingular matrix such that

$$V^{-1}x=\beta e_1 \tag{3.16}$$

for some β. Then $V^{-1}\tilde{C}_{SW}$ has nonzero elements only in its first row. Thus if we apply $\operatorname{diag}\{I_k,V^{-1},1\}$ to the left-hand side of (3.15), we will obtain a pencil of the form

$$\begin{bmatrix} C_{11} & * \\ 0 & * \end{bmatrix} - \mu \begin{bmatrix} N_{k+1} & * \\ 0 & * \end{bmatrix},$$

where C_{11} is $(k+1)\times(k+1)$. We now complete the transformation by applying $\operatorname{diag}\{I_{k+1},V\}$ on the right to obtain

$$\begin{bmatrix} C_{11} & C_{12} \\ 0 & C_{22} \end{bmatrix} - \mu N_{k+m+2}. \tag{3.17}$$

Since this pencil is regular, both of the subpencils $C_{11}-\mu N_{k+1}$ and $C_{22}-\mu N_{m+1}$ must be regular. It is not hard to see that the eigenvalues of $C_{11}-\mu N_{k+1}$ are $\tau_1,\ldots,\tau_k$, and ∞. We just need to show that (3.14) continues to hold when $\tilde{C}_{NW}$ is replaced by C_{11}. But C_{11} differs from $\tilde{C}_{NW}$ only in the last row, so we just have to check that the last row of the equation remains valid. Let c^T denote the last row of C_{11}. We have to show that $c^T\hat{R}=0$. But c^T is also the first row of $V^{-1}\tilde{C}_{SW}$, so $c^T\hat{R}=e_1^TV^{-1}\tilde{C}_{SW}\hat{R}=e_1^TV^{-1}0=0$, and we are done. Thus $\tau_1,\ldots,\tau_k$ are eigenvalues of $C_{11}-\mu N_{k+1}$. Of course ∞ is also an eigenvalue, since N_{k+1} is singular. Finally, $C_{22}-\mu N_{m+1}$ must carry the complement of the spectrum, namely, $\sigma_1,\ldots,\sigma_m$ and ∞. Our mission is accomplished.

Summary of the procedure. Practical matters. The transformation from (3.1) to (3.2) is a product of three transformations involving matrices U, G, and V satisfying (3.7), (3.12), and (3.16), respectively. U and V accomplish simple tasks that can be handled by a reflector, a Gaussian elimination transformation, or some

other simple rank-one modification of I. G is a matrix that accomplishes a GR decomposition of $W \in \mathbb{C}^{(m+k)\times k}$, so it will normally be a product of k such simple transformations. The minimal requirement on all of these matrices is that they be nonsingular, but we would also like them to be well conditioned for stability's sake.

Before we can compute the GR decomposition (3.12), we need to calculate X by solving the Sylvester-like equation (3.10). One approach is to use a variant of the Bartles-Stewart algorithm [8, p. 388], which exploits the structure of the equation. However, this may not be worthwhile if m and k are small. Equation (3.10) is a system of $(m+1)k$ linear equations. If, for example, $m = k = 2$, we have six equations in six unknowns, which can be solved cheaply by standard software.

The function of the matrix V^{-1} is to map each column of $\tilde{C}_{SW}$ to a multiple of e_1. This is possible in principle because the columns of $\tilde{C}_{SW}$ are proportional. In practice, however, the columns will not be exactly proportional because of roundoff errors, so there is no one vector x that is proportional to all of the columns. How does one choose x then? Our solution is to calculate the singular value decomposition of $\tilde{C}_{SW}$. This is inexpensive if k and m are small. We then take x to be the left singular vector corresponding to the maximum singular value. This guarantees that the numbers that this transformation is supposed to make zero are no bigger than $\|V^{-1}\| \sum_{i>1} \sigma_i$ in practice. We have to set these numbers to zero in order to continue the computation. We can do so without compromising backward stability only if the numbers are tiny, i.e., on the level of the machine precision relative to $\|A\|$.

It would be nice if we could prove that these numbers are always tiny, but existing results [1] suggest that this may not be possible. The accuracy of the transformation V depends on how accurately the Sylvester-like equation (3.10) is solved. An accurate solution can be guaranteed if the Sylvester operator $X \to B_{11}X - N_{11}XF$ is well conditioned. We know that this operator is nonsingular if and only if the σ_i are all distinct from the τ_j. Thus it is reasonable to expect that the operator will be well conditioned if the σ_i are well separated from the τ_j. Since we control the shifts, we can always arrange for good separation, which is also desirable from the standpoint of convergence. Unfortunately, good separation of the shifts does not absolutely guarantee a well-conditioned Sylvester operator. Consequently, we cannot guarantee backward stability without actually checking the numbers that are to be set to zero. Conversely, we *can* guarantee backward stability by performing the bulge exchange tentatively and checking the numbers (cf. [1], [10], [9]). If they are not small enough, we refuse to perform the exchange. Instead we can either form one big bulge from the two smaller ones and chase it in one direction or the other or, equivalently, chase one of the bulges back to its point of origin and follow it with the other. If adequate shift separation is maintained, events of this type should be rare.

The bulge swapping procedure is summarized as follows:

- Calculate U satisfying (3.7). Extract Y from U.
- Calculate F using (3.8).
- Solve (3.10) for X.
- Perform right transformation involving U and left transformation involving U^{-1} to get pencil $\tilde{B} - \mu N_{m+k+1}$.
- Calculate G satisfying (3.12).
- Perform left transformation involving G^{-1} and right transformation involving G to get pencil $\tilde{C} - \mu N_{m+k+1}$.
- Calculate singular value decomposition (SVD) of $\tilde{C}_{SW}$. Let x be the dominant left singular vector.

- Use x to construct V satisfying (3.16).
- Perform left transformation involving V^{-1} and right transformation involving V to get (3.17).
- Check that the numbers that were to be zeroed out really are small. If they are small, then set them to zero. Otherwise, refuse to perform the swap.

4. Numerical results. To see how these ideas work in practice, we wrote a Fortran 90 bidirectional double-shift real QR code. This means that $m = k = 2$, the code handles real matrices, and all transformations are real and orthogonal (Householder reflectors). Complex shifts are allowed, but they must occur in conjugate pairs. The purpose of this exercise was simply to find out whether the exchange procedure works. We know of no reason to believe that a bidirectional QR algorithm will be faster or more accurate than a conventional QR code in a serial setting.

The shifting strategy was derived from the standard strategy. That is, σ_1 and σ_2 (resp., τ_1, τ_2) are taken to be the eigenvalues of the lower right-hand (resp., upper left-hand) 2×2 principal submatrix of A. Before an iteration is started, the shifts are checked to make sure they are not too close together. If

$$\min_{i,j} |\, \sigma_i - \tau_j \,| \geq 10^{-3} \max_{i,j} |\, a_{ij} \,|,$$

the iteration is undertaken with the given shifts. Otherwise, the τ_j are perturbed by an ad hoc procedure that guarantees the desired separation.

The shifting strategy is decidedly primitive; surely there are better strategies. For example, one can choose shifts from among the eigenvalues of larger principal submatrices, as suggested in [17].

Each bulge exchange was performed tentatively. After the swap, the part of the matrix that contains the two bulges has the form

$$\left[\begin{array}{ccc|ccc} b_{11} & b_{12} & b_{13} & * & * & * \\ b_{21} & b_{22} & b_{23} & * & * & * \\ b_{31} & b_{32} & b_{33} & * & * & * \\ \hline \delta_{11} & \delta_{12} & \delta_{13} & c_{11} & c_{12} & c_{13} \\ \delta_{21} & \delta_{22} & \delta_{23} & c_{21} & c_{22} & c_{23} \\ 0 & 0 & 0 & c_{31} & c_{32} & c_{33} \end{array}\right],$$

where the δ_{ij} are the numbers that should be zero. The swap is accepted, and the δ_{ij} are set to zero, if

$$\max_{ij} |\, \delta_{ij} \,| \leq 10\epsilon \max \left\{ \max_{ij} |\, b_{ij} \,|, \max_{ij} |\, c_{ij} \,| \right\}, \tag{4.1}$$

where ϵ is the machine precision. If the swap is rejected, the bulge that came up from below is chased back to the bottom, after which the bulge that came from above is also chased to the bottom.[7]

This test is more stringent than is necessary for normwise backward stability; a threshold of $10\epsilon \| A \|$ would be good enough. We chose the stricter criterion in order to improve the accuracy of small eigenvalues of graded matrices.

[7] The tentative swaps are performed as follows: The 6×6 submatrix that contains the two bulges is copied into a small scratch array, and the swap is carried out there. If the swap is rejected, the scratch work is simply discarded. Thus it is never necessary to "undo" a rejected swap.

The code was tested on numerous upper Hessenberg matrices, including the following seven examples. Matrix 1 is an 800×800 matrix created by filling an array with random numbers (normal with mean 0 and variance 1) and reducing it to upper Hessenberg form.

Matrix 2 is a 700×700 matrix with known eigenvalues, which was created as follows. A quasi-triangular matrix with 50 real and 650 complex (random) eigenvalues and a modest departure from normality was built. Then a random orthogonal similarity transformation was applied. Finally, the matrix was reduced to upper Hessenberg form.

Matrix 3, also 700×700, was constructed by the same procedure as matrix 2, except that the departure from normality was made ten times as great. The eigenvalues of this matrix are somewhat ill conditioned.

Matrix 4 is a 625×625 matrix with characteristic polynomial $(x^{125} - 2)^5$, constructed as follows. A random orthogonal similarity transformation was applied to the companion matrix of $(x^{125} - 2)^5$. Then the matrix was reduced to upper Hessenberg form. All eigenvalues have geometric multiplicity 1, algebraic multiplicity 5, and infinitessimal condition number ∞.

Matrix 5 is a 750×750 matrix created by filling the upper Hessenberg part of an array with random numbers.

Matrix 6 is a graded 750×750 matrix obtained by creating an upper Hessenberg matrix like matrix 5, then multiplying the jth row by 1.1^{n-j+1}, $j = 1, \ldots, 750$.

Matrix 7 is a 600×600 matrix with known eigenvalues constructed in the same way as matrix 2, except that the eigenvalues $1.1^m(1 \pm i)$, $m = 1, \ldots, 300$ were assigned.

Computing times and accuracies are recorded in Tables 4.1 and 4.2, respectively. All computations were performed in IEEE standard double precision arithmetic on a DEC AlphaStation 500/333. Four different methods were used. The method denoted "BIQR" is the bidirectional QR algorithm. The method "QR" denotes the same Fortran 90 code as BIQR with the parameters set so that bulges are chased downward only. Thus it is a standard double-shift QR code. Similarly, "RQ" is the same code set so that it chases bulges upward only; it is a standard double-shift RQ code. "LAPACK" denotes the LAPACK [6] code DHSEQR, which performs the multishift QR algorithm. The BIQR, QR, and RQ codes are written in vanilla Fortran 90; we have not attempted to make them fast. In contrast, the LAPACK code was run using tuned level-2 BLAS from DEC's DXML library. Thus it is significantly faster in most cases. If one wishes to compare only comparable codes, one should compare BIQR to QR and RQ. Table 4.1 shows that BIQR is about as fast as QR and RQ in most cases. Similar results were obtained for smaller matrices.

TABLE 4.1
Time in seconds to calculate eigenvalues.

		Method			
Matrix	n	BIQR	QR	RQ	LAPACK
1	800	56	56	61	32
2	700	37	35	38	20
3	700	39	37	40	21
4	625	20	18	19	13
5	750	51	50	48	28
6	750	14	07	29	11
7	600	21	12	14	14

For those matrices whose eigenvalues are known, each number listed in Table 4.2

is the maximum relative error over all eigenvalues of the matrix. For matrices whose eigenvalues are not known, the values computed by the LAPACK code were taken to be the "true" eigenvalues. Table 4.2 shows that BIQR is as accurate as any of the other methods.

TABLE 4.2
Maximum relative error in computed eigenvalues.

	Method			
Matrix	BIQR	QR	RQ	LAPACK
1	4×10^{-14}	4×10^{-14}	3×10^{-14}	———
2	9×10^{-14}	5×10^{-14}	3×10^{-14}	4×10^{-14}
3	1×10^{-03}	1×10^{-03}	1×10^{-03}	1×10^{-03}
4	6×10^{-05}	6×10^{-05}	6×10^{-05}	6×10^{-05}
5	3×10^{-12}	3×10^{-12}	3×10^{-12}	———
6	9×10^{-13}	8×10^{-13}	$5 \times 10^{+01}$	———
7	3×10^{-05}	3×10^{-05}	3×10^{-05}	3×10^{-05}

BIQR did not reject any bulge swaps on Matrices 1–5. About 4000 iterations were performed altogether in these five computations. Numerous additional tests on matrices of these types showed that bulge exchange rejections are extremely rare.

The results for matrix 6, the graded matrix, are quite interesting. All methods ran much faster on this matrix than on the ungraded matrix 5. The considerable savings were due to the matrix's tendency to split apart during QR iterations. Notice that in this case "QR" is significantly faster than LAPACK. Most methods were able to resolve even the smallest eigenvalues. The exception was RQ, which chases bulges in the "wrong" direction and thereby deflates the large eigenvalues first. Not only did it fail to resolve the small eigenvalues; it was also much slower than the other methods. If the grading is reversed, RQ performs well and QR looks bad (and so does LAPACK).

While computing the eigenvalues of matrix 6, BIQR rejected 62 bulge swaps during iterations 67–132. Many of these just barely failed to pass the test (4.1). Every one of them would have passed the less stringent test $\max_{ij} |\delta_{ij}| < 10\epsilon \|A\|$, but then the smallest eigenvalues would not have been resolved as well. When the run was repeated with the test (4.1) turned off, the maximum relative error jumped from 10^{-12} to 10^{-8}.

On matrix 7, BIQR rejected 59 bulge swaps in the first 66 iterations. Turning off the test (4.1) caused no loss of accuracy in this case.

These experiments do not demonstrate the superiority of the code BIQR; indeed, they were not expected to. They do demonstrate the viability of the bulge swapping procedure, which has several potential applications to eigenvalue computations.

REFERENCES

[1] Z. Bai and J. W. Demmel, *On swapping diagonal blocks in real Schur form*, Linear Algebra Appl., 186 (1993), pp. 73–95.

[2] R. Byers, *Hamiltonian and Symplectic Algorithms for the Algebraic Riccati Equation*, Ph.D. thesis, Cornell University, Ithaca, NY, 1983.

[3] R. Byers, *A Hamiltonian QR algorithm*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 212–229.

[4] K.-W. E. Chu, *The solution of the matrix equations* $AXB - CXD = E$ *and* $(YA - DZ, YC - BZ) = (E, F)$, Linear Algebra Appl., 93 (1987), pp. 93–105.

[5] J. J. Dongarra, G. A. Geist, and C. H. Romine, *Fortran subroutines for computing the eigenvalues and eigenvectors of a general matrix by reduction to general tridiagonal form*, ACM Trans. Math. Software, 18 (1992), pp. 392–400.

[6] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, A. McKenney, S. Ostrouchov, and D. Sorensen, *LAPACK Users' Guide*, SIAM, Philadelphia, 1992.
[7] J. G. F. Francis, *The QR transformation, parts I and II*, Comput. J., 4 (1961), pp. 265–272; 332–345.
[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, third ed., Johns Hopkins University Press, Baltimore, 1996.
[9] B. Kågström and P. Poromaa, *Computing eigenspaces with specified eigenvalues of a regular matrix pair* (A, B) *and condition estimation: Theory, algorithms, software*, Numer. Algorithms, 12 (1996), pp. 369–407.
[10] B. Kågström and P. Poromaa, *Lapack-style algorithms and software for solving the generalized Sylvester equation and estimating the separation between regular matrix pairs*, LAPACK Working Note 75, ACM Trans. Math. Software, 22 (1996), pp. 78–103.
[11] C. B. Moler and G. W. Stewart, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241–256.
[12] H. Rutishauser, *Solution of eigenvalue problems with the LR-transformation*, Nat. Bur. Stand. Appl. Math. Ser., 49 (1958), pp. 47–81.
[13] G. W. Stewart, *On the sensitivity of the eigenvalue problem* $Ax = \lambda Bx$, SIAM J. Numer. Anal., 9 (1972), pp. 669–686.
[14] G. W. Stewart, *Error and perturbation bounds for subspaces associated with certain eigenvalue problems*, SIAM Rev., 15 (1973), pp. 727–764.
[15] D. S. Watkins, *Fundamentals of Matrix Computations*, John Wiley and Sons, New York, 1991.
[16] D. S. Watkins, *Bi-directional chasing algorithms for the eigenvalue problem*, SIAM J. Matrix Anal. Appl., 14 (1993), pp. 166–179.
[17] D. S. Watkins, *Shifting strategies for the parallel QR algorithm*, SIAM J. Sci. Comput., 15 (1994), pp. 953–958.
[18] D. S. Watkins, *The transmission of shifts and shift blurring in the QR algorithm*, Linear Algebra Appl., 241/243 (1996), pp. 877–896.
[19] D. S. Watkins and L. Elsner, *Chasing algorithms for the eigenvalue problem*, SIAM J. Matrix Anal. Appl., 12 (1991), pp. 374–384.
[20] D. S. Watkins and L. Elsner, *Convergence of algorithms of decomposition type for the eigenvalue problem*, Linear Algebra Appl., 143 (1991), pp. 19–47.
[21] D. S. Watkins and L. Elsner, *Theory of decomposition and bulge-chasing algorithms for the generalized eigenvalue problem*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 943–967.
[22] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

SIAM J. MATRIX ANAL. APPL.
Vol. 19, No. 4, pp. 1097–1110, October 1998

016

A MODIFIED CHOLESKY ALGORITHM BASED ON A SYMMETRIC INDEFINITE FACTORIZATION*

SHEUNG HUN CHENG† AND NICHOLAS J. HIGHAM†

Abstract. Given a symmetric and not necessarily positive definite matrix A, a modified Cholesky algorithm computes a Cholesky factorization $P(A+E)P^T = R^TR$, where P is a permutation matrix and E is a perturbation chosen to make $A+E$ positive definite. The aims include producing a small-normed E and making $A+E$ reasonably well conditioned. Modified Cholesky factorizations are widely used in optimization. We propose a new modified Cholesky algorithm based on a symmetric indefinite factorization computed using a new pivoting strategy of Ashcraft, Grimes, and Lewis. We analyze the effectiveness of the algorithm, both in theory and practice, showing that the algorithm is competitive with the existing algorithms of Gill, Murray, and Wright and Schnabel and Eskow. Attractive features of the new algorithm include easy-to-interpret inequalities that explain the extent to which it satisfies its design goals, and the fact that it can be implemented in terms of existing software.

Key words. modified Cholesky factorization, optimization, Newton's method, symmetric indefinite factorization

AMS subject classification. 65F05

PII. S0895479896302898

1. Introduction.

Modified Cholesky factorization is a widely used technique in optimization; it is used for dealing with indefinite Hessians in Newton methods [11], [21] and for computing positive definite preconditioners [6], [20]. Given a symmetric matrix A, a modified Cholesky algorithm produces a symmetric perturbation E such that $A+E$ is positive definite, along with a Cholesky (or $\mathrm{LDL^T}$) factorization of $A+E$. The objectives of a modified Cholesky algorithm can be stated as follows [21].

O1. If A is "sufficiently positive definite" then E should be zero.

O2. If A is indefinite, $\|E\|$ should not be much larger than

$$\min\{\,\|\Delta A\| : A+\Delta A \text{ is positive definite}\,\}$$

for some appropriate norm.

O3. The matrix $A+E$ should be reasonably well conditioned.

O4. The cost of the algorithm should be the same as the cost of standard Cholesky factorization to highest order terms.

Two existing modified Cholesky algorithms are one by Gill, Murray, and Wright [11, section 4.4.2.2], which is a refinement of an earlier algorithm of Gill and Murray [10], and an algorithm by Schnabel and Eskow [21].

The purpose of this work is to propose an alternative modified Cholesky algorithm that has some advantages over the existing algorithms. In outline, our approach is to compute a symmetric indefinite factorization

$$PAP^T = LDL^T, \tag{1.1}$$

*Received by the editors April 26, 1996; accepted for publication (in revised form) by P. Gill June 4, 1997; published electronically July 17, 1998. The research of the second author was supported by Engineering and Physical Sciences Research Council grant GR/H/94528.

http://www.siam.org/journals/simax/19-4/30289.html

†Department of Mathematics, University of Manchester, Manchester, M13 9PL, England (chengsh@ma.man.ac.uk, na.nhigham@na-net.ornl.gov).

where P is a permutation matrix, L is unit lower triangular, and D is block diagonal with diagonal blocks of dimension 1 or 2, and to provide the factorization

$$P(A+E)P^T = L(D+F)L^T, \tag{1.2}$$

where F is chosen so that $D+F$ (and hence also $A+E$) is positive definite.[1] This approach is not new; it was suggested by Moré and Sorensen [19] for use with factorizations (1.1) computed with the Bunch–Kaufman [3] and Bunch–Parlett [4] pivoting strategies. However, for neither of these pivoting strategies are all the conditions (O1)–(O4) satisfied, as is recognized in [19]. The Bunch–Parlett pivoting strategy requires $O(n^3)$ comparisons for an $n \times n$ matrix, so condition (O4) does not hold. For the Bunch–Kaufman strategy, which requires only $O(n^2)$ comparisons, it is difficult to satisfy conditions (O1)–(O3), as we explain in section 3.

We use a new pivoting strategy for the symmetric indefinite factorization devised by Ashcraft, Grimes, and Lewis [2], for which conditions (O1)–(O3) are satisfied to within factors depending only on n and for which the cost of the pivot searches is *usually* negligible. We describe this so-called bounded Bunch–Kaufman (BBK) pivoting strategy and its properties in the next section.

There are two reasons why our algorithm might be preferred to those of Gill, Murray, and Wright and of Schnabel and Eskow (henceforth denoted the GMW algorithm and the SE algorithm, respectively). The first is a pragmatic one: we can make use of any available implementation of the symmetric indefinite factorization with the BBK pivoting strategy, needing to add just a small amount of post-processing code to form the modified Cholesky factorization. In particular, we can use the efficient implementations for both dense and sparse matrices written by Ashcraft, Grimes, and Lewis [2], which make extensive use of levels 2 and 3 BLAS for efficiency on high-performance machines. In contrast, in coding the GMW and SE algorithms one must either begin from scratch or make nontrivial changes to an existing Cholesky factorization code.

The second attraction of our approach is that we have a priori bounds that explain the extent to which conditions (O1)–(O3) are satisfied—essentially, if L is well conditioned then an excellent modified Cholesky factorization is guaranteed. For the GMW and SE algorithms it is difficult to describe under what circumstances the algorithms can be guaranteed to perform well.

2. Pivoting strategies. We are interested in symmetric indefinite factorizations (1.1) computed in the following way. If the symmetric matrix $A \in \mathbb{R}^{n\times n}$ is nonzero, we can find a permutation Π and an integer $s = 1$ or 2 so that

$$\Pi A \Pi^T = \begin{array}{c} \\ s \\ n-s \end{array} \begin{array}{c} \begin{array}{cc} s & n-s \end{array} \\ \begin{bmatrix} E & C^T \\ C & B \end{bmatrix} \end{array},$$

with E nonsingular. Having chosen such a Π we can factorize

$$\Pi A \Pi^T = \begin{bmatrix} I_s & 0 \\ CE^{-1} & I_{n-s} \end{bmatrix} \begin{bmatrix} E & 0 \\ 0 & B - CE^{-1}C^T \end{bmatrix} \begin{bmatrix} I_s & E^{-1}C^T \\ 0 & I_{n-s} \end{bmatrix}. \tag{2.1}$$

[1] Strictly, (1.2) is not a Cholesky factorization, since we allow $D+F$ to have 2×2 diagonal blocks, but since any such blocks are positive definite it seems reasonable to use the term "modified Cholesky factorization."

This process is repeated recursively on the $(n-s) \times (n-s)$ Schur complement

$$S = B - CE^{-1}C^T,$$

yielding the factorization (1.1) on completion. This factorization costs $n^3/3$ operations (the same cost as Cholesky factorization of a positive definite matrix) plus the cost of determining the permutations Π.

The Bunch–Parlett pivoting strategy [4] searches the whole submatrix S at each stage, requiring a total of $O(n^3)$ comparisons, and it yields a matrix L whose maximum element is bounded by 2.781. The Bunch–Kaufman pivoting strategy [3], which is used with the symmetric indefinite factorization in both LAPACK [1] and LINPACK [7], searches at most two columns of S at each stage, so it requires only $O(n^2)$ comparisons in total. The Bunch–Kaufman pivoting strategy yields a backward stable factorization [16], but $\|L\|_\infty$ is unbounded, even relative to $\|A\|_\infty$, which makes this pivoting strategy unsuitable for use in a modified Cholesky algorithm, for reasons explained in section 3.

To describe the BBK pivoting strategy [2] it suffices to describe the pivot choice for the first stage of the factorization.

ALGORITHM BBK (BBK pivoting strategy). *This algorithm determines the pivot for the first stage of the symmetric indefinite factorization applied to a symmetric matrix* $A \in \mathbb{R}^{n\times n}$.

$\alpha := (1+\sqrt{17})/8$ (≈ 0.64)
$\gamma_1 :=$ maximum magnitude of any subdiagonal entry in column 1.
If $\gamma_1 = 0$ there is nothing to do on this stage of the factorization.
if $|a_{11}| \geq \alpha\gamma_1$
    use a_{11} as a 1×1 pivot ($s = 1$, $\Pi = I$).
else
    $i := 1$; $\gamma_i := \gamma_1$
    repeat
        $r :=$ row index of first (subdiagonal) entry of maximum magnitude in column i.
        $\gamma_r :=$ maximum magnitude of any off-diagonal entry in column r.
        if $|a_{rr}| \geq \alpha\gamma_r$
            use a_{rr} as a 1×1 pivot ($s = 1$, Π swaps rows and columns 1 and r).
        else if $\gamma_i = \gamma_r$
            use $\begin{bmatrix} a_{ii} & a_{ri} \\ a_{ri} & a_{rr} \end{bmatrix}$ as a 2×2 pivot ($s = 2$, Π swaps rows and columns 1 and i, and 2 and r).
        else
            $i := r$, $\gamma_i := \gamma_r$.
        end
    until a pivot is chosen
end

The repeat loop in Algorithm BBK searches for an off-diagonal element a_{ri} that is simultaneously the largest in magnitude in the rth row and the ith column, and it uses this element to build a 2×2 pivot; the search terminates prematurely if a suitable 1×1 pivot is found.

The following properties noted in [2] are readily verified, using the property that

any 2×2 pivot satisfies

$$\left| \begin{bmatrix} a_{ii} & a_{ri} \\ a_{ri} & a_{rr} \end{bmatrix}^{-1} \right| \leq \frac{1}{\gamma_r(1-\alpha^2)} \begin{bmatrix} \alpha & 1 \\ 1 & \alpha \end{bmatrix}.$$

1. Every entry of L is bounded by $\max\{1/(1-\alpha), 1/\alpha\} \approx 2.78$.
2. Every 2×2 pivot block D_{ii} satisfies $\kappa_2(D_{ii}) \leq (1+\alpha)/(1-\alpha) \approx 4.56$.
3. The growth factor for the factorization, defined in the same way as for Gaussian elimination, is bounded in the same way as for the Bunch–Kaufman pivoting strategy, namely, by $(1+\alpha^{-1})^{n-1} \approx (2.57)^{n-1}$.

Since the value of γ_i increases strictly from one pivot step to the next, the search in Algorithm BBK takes at most n steps. The cost of the searching is intermediate between the cost for the Bunch–Kaufman strategy and that for the Bunch–Parlett strategy. Matrices are known for which the entire remaining submatrix must be searched at each step, in which case the cost is the same as for the Bunch–Parlett strategy. However, Ashcraft, Grimes, and Lewis [2] found in their numerical experiments that on average less than $2.5k$ comparisons were required to find a pivot from a $k \times k$ submatrix, and they give a probabilistic analysis which shows that the expected number of comparisons is less than $ek \approx 2.718k$ for matrices with independently distributed random elements. Therefore we regard the symmetric indefinite factorization with the BBK pivoting strategy as being of similar cost to Cholesky factorization, while recognizing that in certain rare cases the searching overhead may increase the operation count by about 50%.

The symmetric indefinite factorization with the BBK pivoting strategy is backward stable; the same rounding error analysis as for the Bunch–Kaufman pivoting strategy is applicable [2], [16].

The modified Cholesky algorithm of the next section and the corresponding analysis are not tied exclusively to the BBK pivoting strategy. We could use instead the "fast Bunch–Parlett" pivoting strategy from [2], which appears to be more efficient than the BBK strategy when both are implemented in block form [2]. We mention in passing that a block implementation of the SE algorithm has been developed by Daydé [5]. Alternatively, we could use one of the pivoting strategies from [8], [9].

3. The modified Cholesky algorithm. We begin by defining the distance from a symmetric matrix $A \in \mathbb{R}^{n\times n}$ to the symmetric matrices with minimum eigenvalue $\lambda_{\min}$ at least δ, where $\delta \geq 0$:

$$\mu(A,\delta) = \min\{\, \|\Delta A\| : \lambda_{\min}(A+\Delta A) \geq \delta \,\}. \tag{3.1}$$

The distances in the 2- and Frobenius norms, and perturbations that achieve them, are easily evaluated (cf. [12, Thms. 2.1, 3.1]).

THEOREM 3.1. *Let the symmetric matrix $A \in \mathbb{R}^{n\times n}$ have the spectral decomposition $A = Q\Lambda Q^T$ (Q orthogonal, $\Lambda = \operatorname{diag}(\lambda_i)$). Then, for the Frobenius norm,*

$$\mu_F(A,\delta) = \left(\sum_{\lambda_i < \delta} (\delta - \lambda_i)^2 \right)^{1/2}$$

and there is a unique optimal perturbation in (3.1)*, given by*

$$\Delta A = Q \operatorname{diag}(\tau_i) Q^T, \qquad \tau_i = \begin{cases} 0, & \lambda_i \geq \delta, \\ \delta - \lambda_i, & \lambda_i < \delta. \end{cases} \tag{3.2}$$

For the 2*-norm,*

$$\mu_2(A,\delta) = \max(\,0, \delta - \lambda_{\min}(A)\,),$$

and an optimal perturbation is $\Delta A = \mu_2(A,\delta)I$. *The Frobenius norm perturbation* (3.2) *is also optimal in the* 2*-norm.* □

Our modified Cholesky algorithm has a parameter $\delta \geq 0$ and it attempts to produce the perturbation (3.2).

ALGORITHM MC (modified Cholesky factorization). *Given a symmetric matrix* $A \in \mathbb{R}^{n\times n}$ *and a parameter* $\delta \geq 0$ *this algorithm computes a permutation matrix* P, *a unit lower triangular matrix* L, *and a block diagonal matrix* D *with diagonal blocks of dimension* 1 *or* 2 *such that*

$$P(A+E)P^T = LDL^T$$

and $A+E$ *is symmetric positive definite* (*or symmetric positive semidefinite if* $\delta = 0$). *The algorithm attempts to ensure that if* $\lambda_{\min}(A) < \delta$ *then* $\lambda_{\min}(A+E) \approx \delta$.

1. Compute the symmetric indefinite factorization $PAP^T = L\widetilde{D}L^T$ using the BBK pivoting strategy.
2. Let $D = \widetilde{D} + \Delta\widetilde{D}$, where $\Delta\widetilde{D}$ is the minimum Frobenius norm perturbation that achieves $\lambda_{\min}(\widetilde{D} + \Delta\widetilde{D}) \geq \delta$ (thus $\Delta\widetilde{D} = \mathrm{diag}(\Delta\widetilde{D}_{ii})$, where $\Delta\widetilde{D}_{ii}$ is the minimum Frobenius norm perturbation that achieves $\lambda_{\min}(\widetilde{D}_{ii} + \Delta\widetilde{D}_{ii}) \geq \delta$).

To what extent does Algorithm MC achieve the objectives (O1)–(O4) listed in section 1? Objective (O4) is clearly satisfied, provided that the pivoting strategy does not require a large amount of searching, since the cost of step 2 is negligible. For objectives (O1)–(O3) to be satisfied we need the eigenvalues of A to be reasonably well approximated by those of $\widetilde{D}$. For the Bunch–Kaufman pivoting strategy the elements of L are unbounded and the eigenvalues of $\widetilde{D}$ can differ greatly from those of A (subject to A and $\widetilde{D}$ having the same inertia), as is easily shown by example. This is the essential reason why the Bunch–Kaufman pivoting strategy is unsuitable for use in a modified Cholesky algorithm.

To investigate objectives (O1)–(O3) we will make use of a theorem of Ostrowski [18, p. 224]. Here, the eigenvalues of a symmetric $n\times n$ matrix are ordered $\lambda_n \leq \cdots \leq \lambda_1$.

THEOREM 3.2 (Ostrowski). *Let* $M \in \mathbb{R}^{n\times n}$ *be symmetric and* $S \in \mathbb{R}^{n\times n}$ *nonsingular. Then for each* $k = 1\colon n$

$$\lambda_k(SMS^T) = \theta_k\lambda_k(M),$$

where $\lambda_n(SS^T) \leq \theta_k \leq \lambda_1(SS^T)$. □

Assuming first that $\lambda_{\min}(A) > 0$ and applying the theorem with $M = \widetilde{D}$ and $S = L$, we obtain

$$\lambda_{\min}(A) \leq \lambda_{\max}(LL^T)\lambda_{\min}(\widetilde{D}).$$

Now E will be zero if $\lambda_{\min}(\widetilde{D}) \geq \delta$, which is certainly true if

$$\lambda_{\min}(A) \geq \delta\,\lambda_{\max}(LL^T). \tag{3.3}$$

Next, we assume that $\lambda_{\min}(A)$ is negative and apply Theorems 3.1 and 3.2 to obtain

$$\lambda_{\max}(\Delta\widetilde{D}) = \delta - \lambda_{\min}(\widetilde{D}) \leq \delta - \frac{\lambda_{\min}(A)}{\lambda_{\min}(LL^T)}. \tag{3.4}$$

Using Theorem 3.2 again, with (3.4), yields

$$\begin{aligned}\|E\|_2 = \lambda_{\max}(E) &= \lambda_{\max}(L\Delta\widetilde{D}L^T)\\ &\le \lambda_{\max}(LL^T)\lambda_{\max}(\Delta\widetilde{D})\\ &\le \lambda_{\max}(LL^T)\left(\delta - \frac{\lambda_{\min}(A)}{\lambda_{\min}(LL^T)}\right) \qquad (\lambda_{\min}(A) < 0).\end{aligned} \tag{3.5}$$

A final invocation of Theorem 3.2 gives

$$\lambda_{\min}(A+E) \ge \lambda_{\min}(LL^T)\lambda_{\min}(\widetilde{D}+\Delta\widetilde{D}) \ge \lambda_{\min}(LL^T)\delta$$

and

$$\begin{aligned}\|A+E\|_2 = \lambda_{\max}(A+E) &= \lambda_{\max}(L(\widetilde{D}+\Delta\widetilde{D})L^T)\\ &\le \lambda_{\max}(LL^T)\lambda_{\max}(\widetilde{D}+\Delta\widetilde{D})\\ &= \lambda_{\max}(LL^T)\max(\delta, \lambda_{\max}(\widetilde{D}))\\ &\le \lambda_{\max}(LL^T)\max\left(\delta, \frac{\lambda_{\max}(A)}{\lambda_{\min}(LL^T)}\right).\end{aligned}$$

Hence

$$\kappa_2(A+E) \le \kappa_2(LL^T)\max\left(1, \frac{\lambda_{\max}(A)}{\lambda_{\min}(LL^T)\,\delta}\right). \tag{3.6}$$

We can now assess how well objectives (O1)–(O3) are satisfied. To satisfy objective (O1) we would like E to be zero when $\lambda_{\min}(A) \ge \delta$, and to satisfy (O2) we would like $\|E\|_2$ to be not much larger than $\delta - \lambda_{\min}(A)$ when A is not positive definite. The sufficient condition (3.3) for E to be zero and inequality (3.5) show that these conditions do hold modulo factors $\lambda_{\max,\min}(LL^T)$. Inequality (3.6) bounds $\kappa_2(A+E)$ with the expected reciprocal dependence on δ, again with terms $\lambda_{\max,\min}(LL^T)$. The conclusion is that the modified Cholesky algorithm is guaranteed to perform well if $\lambda_{\min}(LL^T)$ and $\lambda_{\max}(LL^T)$ are not too far from 1.

Note that, since L is unit lower triangular, $e_1^T(LL^T)e_1 = 1$, which implies that $\lambda_{\min}(LL^T) \le 1$ and $\lambda_{\max}(LL^T) \ge 1$. For the BBK pivoting strategy we have $\max_{i,j}|l_{ij}| \le 2.781$, so

$$1 \le \lambda_{\max}(LL^T) \le \operatorname{trace}(LL^T) = \|L\|_F^2 \le n + \tfrac{1}{2}n(n-1)2.781^2 \le 4n^2 - 3n. \tag{3.7}$$

Furthermore,

$$1 \le \lambda_{\min}(LL^T)^{-1} = \|(LL^T)^{-1}\|_2 = \|L^{-1}\|_2^2 \le (3.781)^{2n-2}, \tag{3.8}$$

using a bound from [15, Thm. 8.13 and Prob. 8.5]. These upper bounds are approximately attainable, but in practice are rarely approached. In particular, the upper bound of (3.8) can be approached only in the unlikely event that most of the subdiagonal elements of L are negative and of near maximal magnitude. Note that each 2×2 pivot causes a subdiagonal element $l_{i+1,i}$ to be zero and so further reduces the likelihood of $\|L^{-1}\|_2$ being large.

In the analysis above we have exploited the fact that the extent to which the eigenvalues of A and $\widetilde{D}$ agree can be bounded in terms of the condition of L. If L is well conditioned then the singular values of A are close to the moduli of the eigenvalues of $\widetilde{D}$. We are currently exploring the application of this fact to the computation of rank-revealing factorizations.

4. Comparison with the GMW and SE algorithms. The GMW and SE algorithms both carry out the steps of a Cholesky factorization of a symmetric matrix $A \in \mathbb{R}^{n\times n}$, increasing the diagonal entries as necessary in order to ensure that negative pivots are avoided. (Actually, the GMW algorithm works with an $\mathrm{LDL}^{\mathrm{T}}$ factorization, where D is diagonal, but the difference is irrelevant to our discussion.) Hence both algorithms produce Cholesky factors of $P^T(A+E)P$ with a diagonal E. From Theorem 3.1 we note that the "optimal" perturbation in objective (O2) of section 1 is, in general, full for the Frobenius norm and can be taken to be diagonal for the 2-norm (but is generally not unique). There seems to be no particular advantage to making a diagonal perturbation to A. Our algorithm perturbs the whole matrix, in general.

By construction, the GMW and SE algorithms make perturbations E to A that are bounded a priori by functions of n and $\|A\|$ only. The GMW algorithm produces a perturbation E for which

$$\|E\|_\infty \le \left(\frac{\beta}{\xi} + (n-1)\xi\right)^2 + 2(\alpha + (n-1)\xi^2) + \delta, \tag{4.1}$$

where $\delta \ge 0$ is a tolerance,

$$\alpha = \max_i |a_{ii}|, \quad \beta = \max_{i\ne j} |a_{ij}|, \quad \xi^2 = \max\{\,\alpha, \beta/\sqrt{n^2-1}, u\,\},$$

and u is the unit roundoff [11, p. 110]. For the SE algorithm the perturbation is bounded in terms of a certain eigenvalue bound ϕ obtained by applying Gershgorin's theorem:

$$\|E\|_\infty \le \phi + \frac{2\tau}{1-\tau}(\phi + \alpha), \tag{4.2}$$

where τ is a tolerance, suggested in [21] to be chosen as $\tau = u^{1/3}$. The quantity ϕ satisfies $\phi \le n(\alpha+\beta)$, so (4.2) is a smaller bound than (4.1) by about a factor n.

The bounds (4.1) and (4.2) can be compared with (3.5) for Algorithm MC. The bound (3.5) has the advantage of directly comparing the perturbation made by Algorithm MC with the optimal one, as defined by (3.1) and evaluated in Theorem 3.1, and it is potentially a much smaller bound than (4.1) and (4.2) if $|\lambda_{\min}(A)| \ll |\lambda_{\max}(A)|$ and $\kappa_2(LL^T)$ is not too large. On the other hand, the bound (3.5) can be much larger than (4.1) and (4.2) if $\kappa_2(LL^T)$ is large.

All three algorithms satisfy objective (O1) of not modifying a sufficiently positive definite matrix, though for the GMW and SE algorithms no condition analogous to (3.3) that quantifies "sufficiently" in terms of $\lambda_{\min}(A)$ is available. Bounds for $\kappa_2(A+E)$ that are exponential in n hold for the GMW and SE algorithms [21]. The same is true for Algorithm MC: see (3.6)–(3.8).

To summarize, in terms of the objectives of section 1 for a modified Cholesky algorithm, Algorithm MC is theoretically competitive with the GMW and SE algorithms, with the weakness that if $\kappa_2(LL^T)$ is large then the bound on $\|E\|_2$ is weak.

When applied to an indefinite matrix, the GMW and SE algorithms provide information that enables a direction of negative curvature of the matrix to be produced; these directions are required in certain algorithms for unconstrained optimization in order to move away from nonminimizing stationary points. For an indefinite matrix, Algorithm MC provides immediate access to a direction of negative curvature from the

$\mathrm{LDL}^{\mathrm{T}}$ factorization computed in step 1, and because $\kappa(L)$ is bounded, this direction satisfies conditions required for convergence theory [19].

Finally, we consider the behavior of the algorithms in the presence of rounding errors. Algorithm MC is backward stable because the underlying factorization is [2]: barring large element growth in the symmetric indefinite factorization with the BBK pivoting strategy, the algorithm produces $\mathrm{LDL}^{\mathrm{T}}$ factors not of $P(A+E)P^T$, but of $P(A+E+F)P^T$, where $\|F\|_2 \le c_n u \|A+E\|_2$ with c_n a constant. Although no comments on numerical stability are given in [11] and [21], a simple argument shows that the GMW and SE algorithms are backward stable. Apply either algorithm to A, obtaining the Cholesky factorization $P(A+E)P^T = R^TR$. Now apply the same algorithm to $P(A+E)P^T$: it will not need to modify $P(A+E)P^T$, so it will return the same computed R factor. But since no modification was required, the algorithm must have carried out a standard Cholesky factorization. Since Cholesky factorization is a backward stable process, the modified Cholesky algorithm must itself be backward stable.

5. Numerical experiments. We have experimented with MATLAB implementations of Algorithm MC and the GMW and SE algorithms. The M-file for the GMW algorithm was provided by M. Wright and sets the tolerance $\delta = 2u$ (which is the value of MATLAB's variable `eps`). The M-file for the SE algorithm was provided by E. Eskow and sets the tolerance $\tau = (2u)^{1/3}$. In Algorithm MC we set $\delta = \sqrt{u}\|A\|_\infty$.

The aims of the experiments are as follows: to see how well the Frobenius norm of the perturbation E produced by Algorithm MC approximates the distance $\mu_F(A,\delta)$ defined in (3.1), and to compare the norms of the perturbations E and the condition numbers of $A+E$ produced by the three algorithms. We measure the perturbations E by the ratios

$$r_F = \frac{\|E\|_F}{\mu_F(A,\delta)}, \qquad r_2 = \frac{\|E\|_2}{|\lambda_{\min}(A)|},$$

which differ only in their normalization and the choice of norm. Algorithm MC attempts to make r_F close to 1. The quantity r_2 is used by Schnabel and Eskow to compare the performance of the GMW and SE algorithms; since E is diagonal for these algorithms, r_2 compares the amount added to the diagonal with the minimum diagonal perturbation that makes the perturbed matrix positive semidefinite.

First, we note that the experiments of Schnabel and Eskow [21] show that the SE algorithm can produce a substantially smaller value of r_2 than the GMW algorithm. Schnabel and Eskow also identified a 4×4 matrix for which the GMW algorithm significantly outperforms the SE algorithm:

$$A = \begin{bmatrix} 1890.3 & -1705.6 & -315.8 & 3000.3 \\ & 1538.3 & 284.9 & -2706.6 \\ & & 52.5 & -501.2 \\ & & & 4760.8 \end{bmatrix}, \tag{5.1}$$

$$\lambda(A) = \{-0.38, -0.34, -0.25, 8.2 \times 10^3\}.$$

We give results for this matrix in Table 5.1; they show that Algorithm MC can also significantly outperform the SE algorithm.

We ran a set of tests similar to those of Schnabel and Eskow [21]. The matrices A are of the form $A = Q\Lambda Q^T$, where $\Lambda = \mathrm{diag}(\lambda_i)$ with the eigenvalues λ_i from one

TABLE 5.1
Measures of E for 4×4 matrix (5.1).

	MC	GMW	SE
r_F	1.3	2.7	3.7×10^3
r_2	1.7	2.7	2.8×10^3

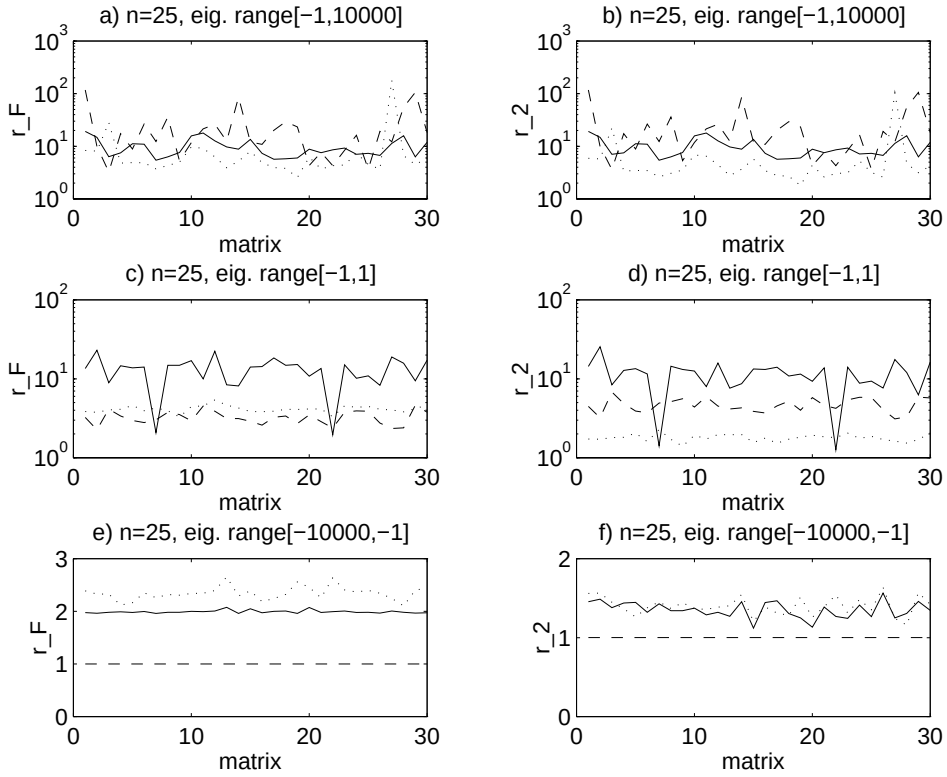


FIG. 5.1. *Measures of E for 30 random indefinite matrices with $n = 25$. Key: GMW —, SE $\cdots$, MC - - -.*

of three random uniform distributions: $[-1, 10^4]$, $[-1, 1]$, and $[-10^4, -1]$. For the first range, one eigenvalue is generated from the range $[-1, 0)$ to ensure that A has at least one negative eigenvalue. The matrix Q is a random orthogonal matrix from the Haar distribution, generated using the routine `qmult` from the Test Matrix Toolbox [14], which implements an algorithm of Stewart [22]. For each eigenvalue distribution we generated 30 different matrices, each corresponding to a fresh sample of Λ and of Q. We took $n = 25, 50, 100$. The ratios r_F and r_2 are plotted in Figures 5.1–5.3. Figure 5.4 plots the condition numbers $\kappa_2(A + E)$ for $n = 25$; the condition numbers for $n = 50$ and $n = 100$ show a very similar behavior. Table 5.2 reports the number of comparisons used by the BBK pivoting strategy on these matrices for each n; the maximum number of comparisons is less than n^2 in each case.

In Figure 5.5 we report results for three nonrandom matrices from the Test Matrix

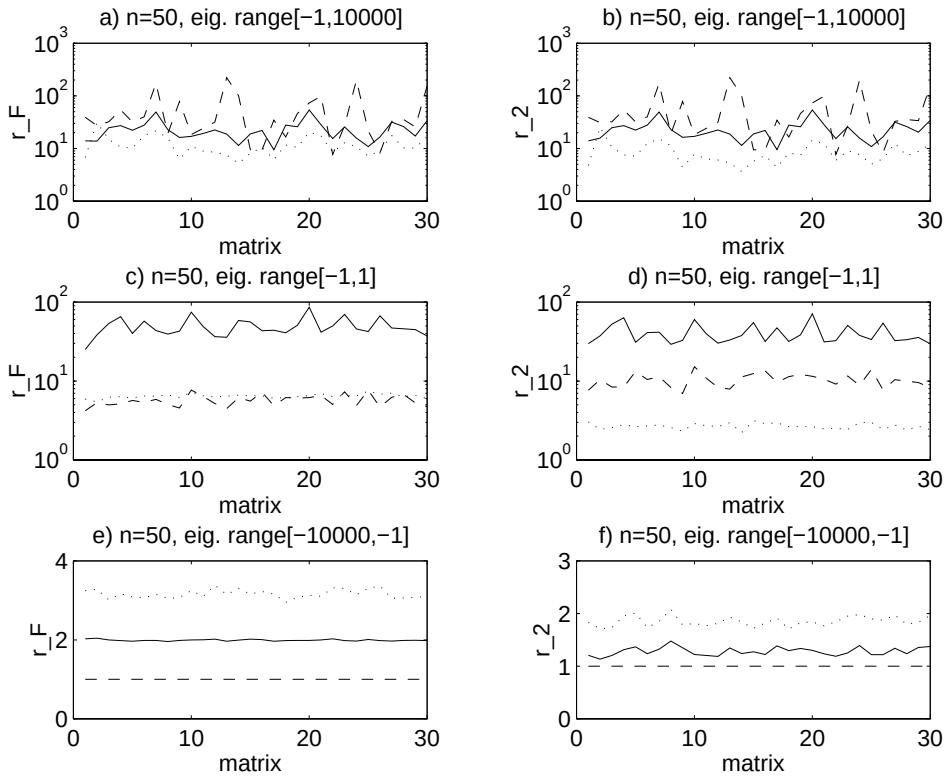


FIG. 5.2. *Measures of E for 30 random indefinite matrices with $n = 50$. Key: GMW —, SE $\cdots$, MC - - -.*

TABLE 5.2
Number of comparisons for BBK pivoting strategy.

n:	25	50	100
max	523	2188	8811
mean	343.9	1432.8	5998.4

Toolbox. `Clement` is a tridiagonal matrix with eigenvalues plus and minus the numbers $n-1, n-3, n-5, \ldots, (1 \text{ or } 0)$. `Dingdong` is the symmetric $n \times n$ Hankel matrix with (i,j) element $0.5/(n-i-j+1.5)$, whose eigenvalues cluster around $\pi/2$ and $-\pi/2$. `Ipjfact` is the Hankel matrix with (i,j) element $1/(i+j)!$.

Our conclusions from the experiments are as follows.

1. None of the three algorithms is uniformly better than the others in terms of producing a small perturbation E, whichever measure r_F or r_2 is used. All three algorithms can produce values of r_F and r_2 significantly greater than 1, depending on the problem.
2. Algorithm MC often achieves its aim of producing $r_F \approx 1$. It produced r_F of order 10^3 for the eigenvalue distribution $[-1, 10^4]$ for each n, and the values of $\kappa_2(LL^T)$ (not shown here) were approximately $100 r_F$ in each such case.

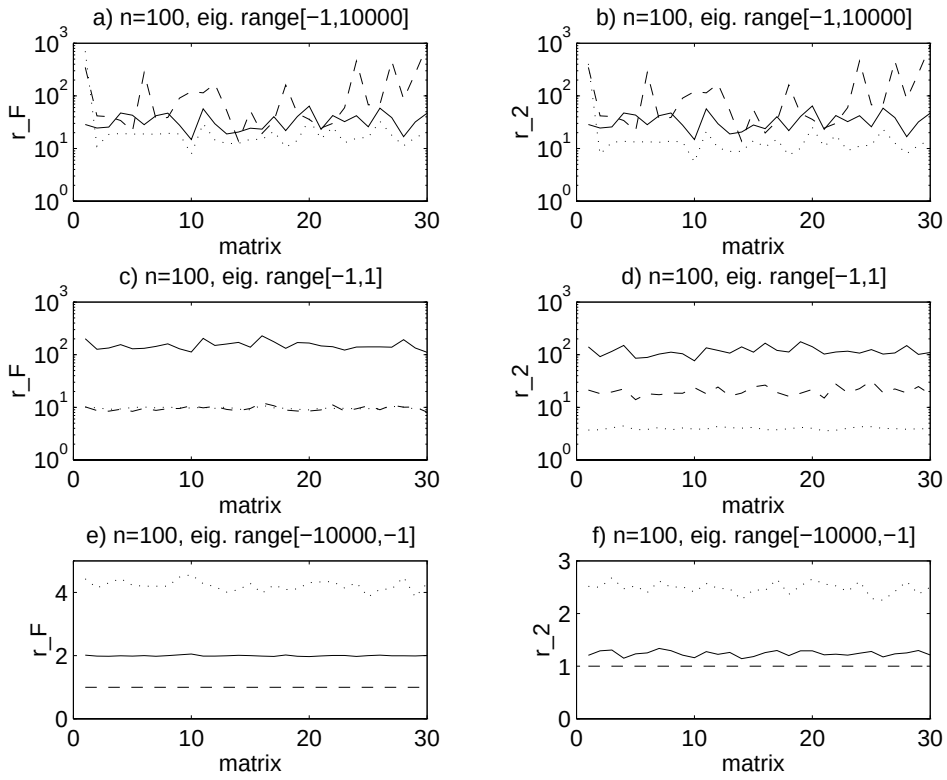


FIG. 5.3. *Measures of E for* 30 *random indefinite matrices with* $n = 100$. *Key: GMW —, SE* $\cdots$, *MC - - -.*

However, often r_F was of order 1 when $\kappa_2(LL^T)$ was of order 10^2 or 10^3, so a large value of $\kappa_2(LL^T)$ is only a necessary condition, not a sufficient one, for poor performance of Algorithm MC; in other words, the bounds of section 3 can be weak.

3. The condition numbers $\kappa_2(A+E)$ vary greatly among the algorithms. Our experience is that for $\delta = \sqrt{u}\|A\|_\infty$ Algorithm MC fairly consistently produces condition numbers of order $100/\sqrt{u}$; the condition number is, as predicted by (3.6), much smaller for the random matrices with eigenvalues on the range $[-10^4, -1]$, because the algorithm attempts to perturb all the eigenvalues to δ. The condition numbers produced by the GMW and SE algorithms vary greatly with the type of matrix.

The fact that r_F is close to 1 for the random matrices with eigenvalues in the range $[-10^4, -1]$ for Algorithm MC is easily explained. Let A be negative definite. Then Algorithm MC computes $P(A + E)P^T = L(\delta I)L^T$. Hence

$$\begin{aligned} r_F &= \frac{\|E\|_F}{(\sum_i (\delta - \lambda_i)^2)^{1/2}} \\ &\le \frac{\|E\|_F}{\|A\|_F} = \frac{\|A - \delta \cdot P^T L L^T P\|_F}{\|A\|_F} \end{aligned}$$

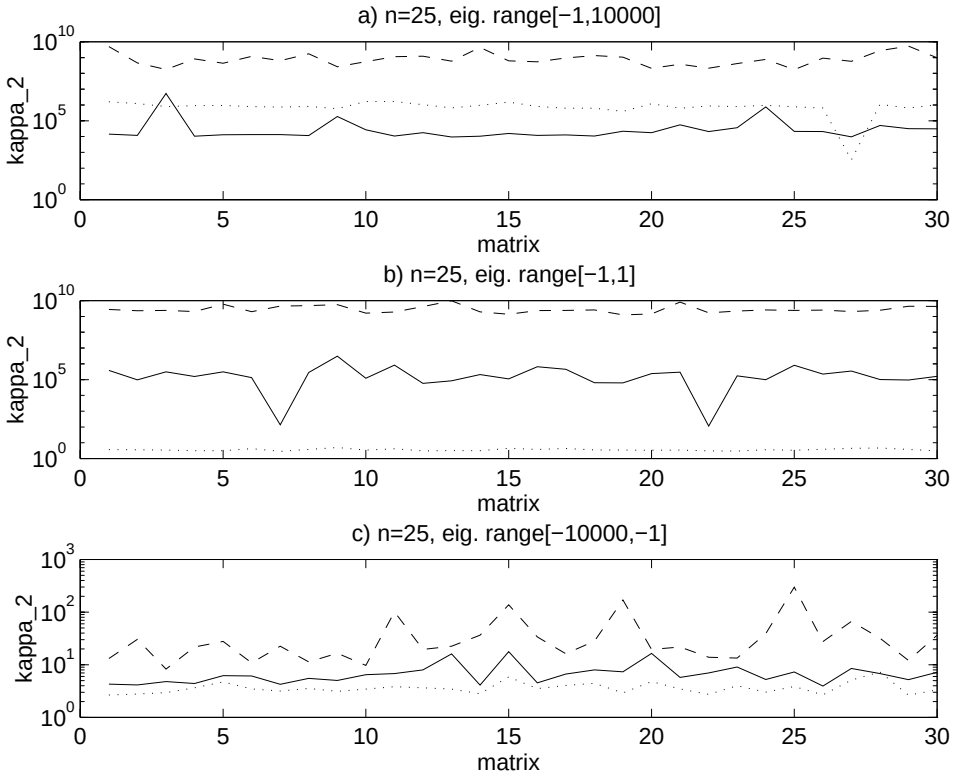


FIG. 5.4. *Condition numbers* $\kappa_2(A+E)$ *for* 30 *random indefinite matrices with* $n = 25$. *Key: GMW —, SE* $\cdots$*, MC - - -.*

$$\leq \frac{\|A\|_F + \delta\|LL^T\|_F}{\|A\|_F}$$
$$\leq 1 + \frac{(4n^2 - 3n)\delta}{\|A\|_F},$$

using (3.7), so r_F can exceed 1 only by a tiny amount for Algorithm MC applied to a negative definite matrix, irrespective of $\kappa_2(LL^T)$.

6. Concluding remarks. Algorithm MC, based on the symmetric indefinite factorization with the bounded Bunch–Kaufman pivoting strategy, merits consideration as an alternative to the algorithms of Gill, Murray, and Wright and Schnabel and Eskow. The results in section 5 suggest that the new algorithm is competitive with the GMW and SE algorithms in terms of the objectives (O1)–(O4) listed in section 1. Algorithm MC has the advantages that the extent to which it satisfies the objectives is neatly, although not sharply, described by the bounds of section 3 and that it can be implemented by augmenting existing software with just a small amount of additional code.

Since all three modified Cholesky algorithms can "fail," that is, they can produce unacceptably large perturbations, it is natural to ask how failure can be detected and what should be done about it. The GMW and SE algorithms produce their (diagonal)

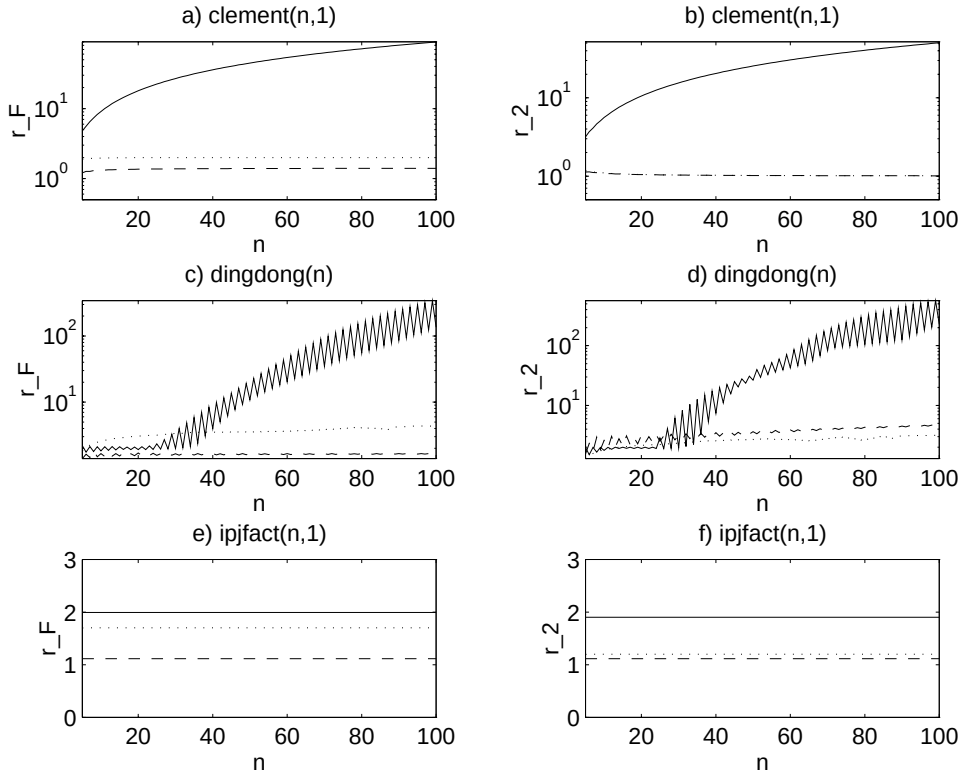


FIG. 5.5. *Measures of E for three nonrandom matrices. Key: GMW —, SE ···, MC - - -.*

perturbations explicitly, so it is trivial to evaluate their norms. For Algorithm MC, the perturbation to A is (see (1.2)) $E = P^T L(D+F)L^T P - A$, which would require $O(n^3)$ operations to form explicitly. However, we can *estimate* $\|E\|_\infty$ using the norm estimator from [13] (which is implemented in LAPACK). The estimator requires the formation of products Ex for certain vectors x, and these can be computed in $O(n^2)$ operations; the estimate produced is a lower bound that is nearly always within a factor 3 of the true norm. For all three algorithms, then, we can inexpensively test whether the perturbation produced is acceptably small. Unfortunately, for none of the algorithms is there an obvious way to improve a modified Cholesky factorization that makes too big a perturbation; whether improvement is possible, preferably cheaply, is an open question. Of course one can always resort to computing an optimal perturbation by computing the eigensystem of A and using the formulae in Theorem 3.1.

We note that we have approached the problem of modified Cholesky factorization from a purely linear algebra perspective. An important test of a modified Cholesky algorithm is to evaluate it in an optimization code on representative problems, as was done by Schlick [20] for the GMW and SE algorithms. This we plan to do for Algorithm MC in future work.

Finally, we mention that a generalization of the modified Cholesky problem motivated by constrained optimization is analyzed in detail in [17].

REFERENCES

[1] E. Anderson, Z. Bai, C. H. Bischof, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, A. McKenney, S. Ostrouchov, and D. C. Sorensen, *LAPACK Users' Guide, Release* 2.0, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1995.

[2] C. Ashcraft, R. G. Grimes, and J. G. Lewis, *Accurate symmetric indefinite linear equation solvers*, SIAM J. Matrix Anal. Appl., to appear.

[3] J. R. Bunch and L. Kaufman, *Some stable methods for calculating inertia and solving symmetric linear systems*, Math. Comp., 31 (1977), pp. 163–179.

[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] M. J. Daydé, *A Block Version of the Eskow–Schnabel Modified Cholesky Factorization*, Technical report RT/APO/95/8, Dept. Informatique et Maths Appls., ENSEEIHT-IRIT, 31071 Toulouse Cedex, France, 1995.

[6] M. J. Daydé, J.-Y. L'Excellent, and N. I. M. Gould, *On the Use of Element-By-Element Preconditioners to Solve Large Scale Partially Separable Optimization Problems*, Report RAL-95-010, Atlas Centre, Rutherford Appleton Laboratory, Didcot, Oxon, UK, 1995.

[7] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, *LINPACK Users' Guide*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1979.

[8] 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.

[9] I. S. Duff, J. K. Reid, N. Munskgaard, and H. B. Nielsen, *Direct solution of sets of linear equations whose matrix is sparse, symmetric and indefinite*, J. Inst. Math. Appl., 23 (1979), pp. 235–250.

[10] P. E. Gill and W. Murray, *Newton-type methods for unconstrained and linearly constrained optimization*, Math. Programming, 7 (1974), pp. 311–350.

[11] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, London, 1981.

[12] N. J. Higham, *Computing a nearest symmetric positive semidefinite matrix*, Linear Algebra Appl., 103 (1988), pp. 103–118.

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

[14] N. J. Higham, *The Test Matrix Toolbox for* Matlab *(Version* 3.0*)*, Numerical Analysis report 276, Manchester Centre for Computational Mathematics, Manchester, England, 1995.

[15] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.

[16] N. J. Higham, *Stability of the diagonal pivoting method with partial pivoting*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 52–65.

[17] N. J. Higham and Sheung Hun Cheng, *Modifying the inertia of matrices arising in optimization*, Linear Algebra Appl., to appear.

[18] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.

[19] J. J. Moré and D. C. Sorensen, *On the use of directions of negative curvature in a modified Newton method*, Math. Programming, 16 (1979), pp. 1–20.

[20] T. Schlick, *Modified Cholesky factorizations for sparse preconditioners*, SIAM J. Sci. Comput., 14 (1993), pp. 424–445.

[21] R. B. Schnabel and E. Eskow, *A new modified Cholesky factorization*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 1136–1158.

[22] G. W. Stewart, *The efficient generation of random orthogonal matrices with an application to condition estimators*, SIAM J. Numer. Anal., 17 (1980), pp. 403–409.